American Psychological Association

Volume 109
Number 6
August 2017

Published eight times

ISSN 0022-0663

Journal of Educational Psychology

www.apa.org/pubs/journals/edu

The main purpose of the *Journal of Educational Psychology*® is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; call 202-336-5600 or 800-374-2721; or visit www.apa.org/pubs/journals/subscriptions.aspx

Manuscripts: Submit manuscripts electronically through the Manuscript Submissions Portal found at www.apa.org/pubs/journals/edu according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Steve Graham, at steve.graham@asu.edu. The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

Electronic Access: APA members who subscribe to this journal have automatic access to all issues of the journal in the PsycARTICLES® full-text database. See http://my.apa.org/access.html.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

APA Journal Staff: Rosemarie Sokol-Chang, PhD, *Publisher, APA Journals*; Mare Meadows, *Managing Director*; Amanda S. Conley, *Journal Production Manager*; Cheryl Johnson, *Editorial Manuscript Coordinator*; Jodi Ashcraft, *Director, Advertising Sales and Exhibits*.

Journal of Educational Psychology® (ISSN 0022-0663) is published eight times (January, February, April, May, July, August, October, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2017 rates follow: *Nonmember Individual:* $250 Domestic, $292 Foreign, $314 Air Mail. *Institutional:* $953 Domestic, $1,030 Foreign, $1,054 Air Mail. *APA Member:* $123. *APA Student Affiliate:* $75. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to *Journal of Educational Psychology*, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Effective with the 1986 volume, this journal is printed on acid-free paper.

Journal of

Educational Psychology®

www.apa.org/pubs/journals/edu

August 2017

Volume 109
Number 6

Reading

Mathematics

Motivation and Schooling

Teacher Judgments

(*Contents continue*)

Other

Journal of Educational Psychology
2017, Vol. 109, No. 6, 741–760

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000168

Web-Based Text Structure Strategy Instruction Improves Seventh Graders' Content Area Reading Comprehension

Kausalai (Kay) Wijekumar
Texas A&M University

Bonnie J. F. Meyer and Puiwa Lei
The Pennsylvania State University

Reading comprehension in the content areas is a challenge for many middle grade students. Text structure-based instruction has yielded positive outcomes in reading comprehension at all grade levels in small and large studies. The text structure strategy delivered via the web, called Intelligent Tutoring System for the Text Structure Strategy (ITSS), has proven successful in large-scale studies at 4th and 5th grades and a smaller study at 7th grade. Text structure-based instruction focuses on selection and encoding of strategic memory. This strategic memory proves to be an effective springboard for many comprehension-based activities such as summarizing, inferring, elaborating, and applying. This was the first large-scale randomized controlled efficacy study on the web-based delivery of the text structure strategy to 7th-grade students. 108 classrooms from rural and suburban schools were randomly assigned to ITSS or control and pretests and posttests were administered at the beginning and end of the school year. Multilevel data analyses were conducted on standardized and researcher designed measures of reading comprehension. Results showed that ITSS classrooms outperformed the control classrooms on all measures with the highest effects reported for number of ideas included in the main idea. Results have practical implications for classroom practices.

Keywords: reading comprehension, intelligent tutoring systems, metacognition, text structure

Reading comprehension is a cornerstone of academic and lifelong learning, economic independence, and civic engagement. Unfortunately, many middle schoolchildren fail to demonstrate mastery of reading comprehension as evidenced by both the National Assessment of Educational Progress (NAEP, 2013) and the Early Childhood Longitudinal Study–Kindergarten Cohort (ECLS-K; Reardon, Valentino, & Shores, 2012). Over 60% of middle schoolchildren score at basic or below basic levels of proficiency in this important foundational skill (NAEP, 2013). The continued poor performance of eighth graders on the NAEP tests shows reading comprehension to be a lingering challenge in the middle grades. Results from the ECLS-K show that students in middle grades have not shown any change in comprehension ability in 30 years and gaps between the socioeconomic groups are widening. The NAEP and ECLS-K results are troubling because over half of middle schoolchildren cannot proficiently learn by reading content curricula and this in turn pushes the comprehension burden into content area classrooms (Allington, 2011; Edmonds et al., 2009). The goal of this project was to intervene at the 7th-grade level to increase the likelihood of the students acquiring reading comprehension skills prior to entering their final year in middle school.

Text structure-based solutions to reading comprehension problems faced by early adolescent readers have shown positive results in upper elementary and middle-grades (Hebert et al., in press; Meyer et al., 2010; Wijekumar, Meyer, & Lei, 2012; Wijekumar et al., 2014). Technology-supported delivery of interventions to children in middle grades has also shown promise for improved outcomes in reading comprehension (Slavin, Cheung, Groff, & Lake, 2008). The Intelligent Tutoring System for the Structure Strategy (ITSS) combines the text structure strategy and web-based technologies to improve content area reading comprehension. ITSS has been tested in large-scale randomized controlled studies with 4th- and 5th-grade children (Wijekumar et al., 2012, 2014). With 7th- grade students, ITSS was examined using a pretest and posttest(s) design study, where students were randomly assigned to variations in ITSS adaptations (e.g., types of feedback; Meyer et al., 2010). All three studies showed positive results favoring students using ITSS. Large-scale and methodologically rigorous randomized controlled trials that are effectively implemented to inform practice are needed in order to draw any causal conclusions about the ITSS intervention with 7th graders.

This article describes one such efficacy study that sought to strengthen the research base on improving 7th-grade students' reading comprehension by reporting on a recent large-scale multisite randomized controlled trial with 108 rural and suburban

This article was published Online First March 13, 2017.

Kausalai (Kay) Wijekumar, Department of Teaching, Learning and Culture, Director, Center for Urban School Partnerships, Texas A&M University; Bonnie J. F. Meyer and Puiwa Lei, Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A080133 to The Pennsylvania State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Kausalai (Kay) Wijekumar, Department of Teaching, Learning and Culture, Center for Urban School Partnerships, Texas A&M University, 420C Harrington Tower, 4232 TAMU, College Station, TX 77843. E-mail: K_Wijekumar@tamu.edu

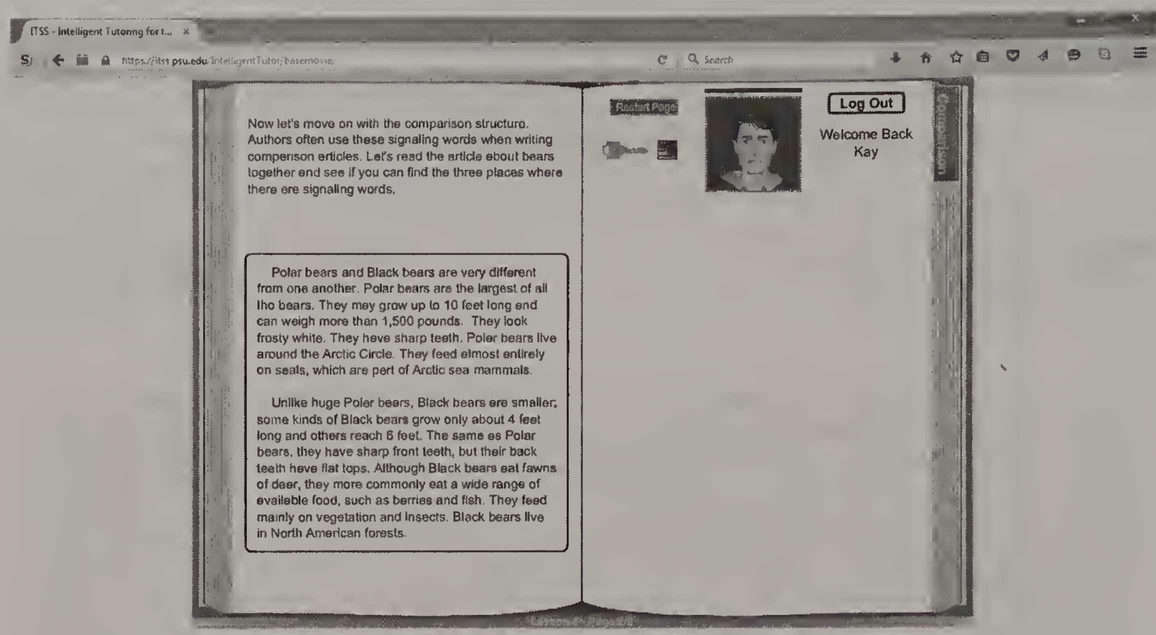


Figure 1. Web-based ITSS interface.

classrooms. ITSS uses a web-based interface shown in Figure 1 to deliver text structure intervention for students, teaching them how to create strategic memory about expository text. ITSS is designed to be delivered as a partial substitute to the language arts curriculum once a week for about 30–45 min, contains over 100 lessons, and focuses on expository texts from science, social studies, current events, and sports with readability from 2nd-grade to 12th-grade levels. This study was powered to answer one primary confirmatory research question about reading comprehension outcomes using a standardized distal test and researcher designed proximal and distal measures. We conducted further analyses to answer five exploratory questions to see whether effects of intervention vary by gender, prior knowledge, and locale that are of interest to researchers and practitioners.

Causes for Reading Comprehension Problems and Potential Remedy

Comprehending content area texts is a difficult task that requires students to fluidly bootstrap a complex set of cognitive and metacognitive skills. Cognitive skills for middle-grade learners who need to read and comprehend content area texts include vocabulary knowledge, contextual and background knowledge, linguistic awareness, and strategies to select and encode ideas into memory structures that are well-associated, integrated with prior knowledge, and available for use in a multitude of activities (e.g., problem solving, writing, inferring, and elaborating; Kintsch, 1998; Pressley, 2000; van Dijk & Kintsch, 1983). Metacognitive skills include awareness of and appropriate use of strategies, planning and monitoring of the comprehension process, and effective allocation of mental resources to the task at hand (Schellings & Broekkamp, 2011).

There are three possible causes for the reading comprehension deficits facing middle-grade students relating to the reader, text, and task (Meyer, 1975). First, a lack of comprehension component skills to master the content may hinder the *reader*'s comprehension. Research has shown that children with poor comprehension skills engage in little to no planning before, during, or after reading (Mason, 2004). They are unable to identify important elements of the text, summarize the text effectively, construct strategically organized recalls showing strong cohesion (Meyer, Brandt, & Bluth, 1980), make inferences, and/or integrate their prior knowledge with the new information (McNamara, 2004). Students also lack the ability to monitor their comprehension and exhibit low motivation and efficacy toward reading (Taboada, Tonks, Wigfield, & Guthrie, 2009).

A second possible cause for reading comprehension challenges facing middle-grade learners is novel or complex content, which requires effortful processing that is unfamiliar or unpracticed by the learners (Meyer et al., 1980). Texts for middle-grade learners are often difficult to read and understand because they are complex and dense, low in cohesion requiring inferences and elaboration, and structurally unfamiliar (Caccamise, Friend, Groneman, Littrell-Baez, & Kintsch, 2014; Duke, 2010; Kintsch & Kintsch, 2004). Middle-grade content may be uninteresting to the students and offer few choices to motivate and maintain interest in continued focused reading activities (Guthrie & Davis, 2003; Taboada et al., 2009). There is also an assumption that middle-grade learners will have strong prior knowledge allowing them to read, understand, and connect to new information gathered from reading.

Finally, concern has been expressed by researchers about the lack of a strong research base to support the instructional *tasks* (Slavin et al., 2008) included in textbooks and/or instructional materials and delivered by knowledgeable teachers at the middle-school level (Allington, 2011; Edmonds et al., 2009). A recent review by Wijekumar, Meyer, Harris, Graham, and Beerwinkle (2016) shows how language arts instruction varies greatly based on the strategies taught and sometimes contradicts proven practices. Content area teachers often assume that students have completed the learning to read phase of reading instruction in elementary school, and, therefore, do not need additional instruction about comprehension in middle school (Allington, 2011; Pressley, 2000; Raudenbush, Rowan, & Cheong, 1993). Additionally, middle-grade teachers are not typically trained to deliver comprehension-related instruction in a content area classroom (Allington, 2011).

The focus of this study is to address these possible causes for reading comprehension challenges with 7th-grade learners by teaching them how to select and encode coherent strategic memory of text using five text structures: comparison, problem and solution, cause and effect, sequence, and description and nested text structures (e.g., comparison of solutions within the problem and solution text structure, Meyer, 1975; Meyer & Wijekumar, 2007; Wijekumar et al., 2012). We focus on 7th-grade learners because of the importance of addressing reading comprehension difficulties with students prior to entering their final year in middle school. The academic rigor and complexity increases as students move to 8th grade and beyond to high school. Preparing the students at 7th grade may alleviate potential challenges as they move forward.

The second focus is on instruction framed by the five text structures to teach students to be strategic in reading and comprehending content area texts, support students with strong or weak prior knowledge, and provide a useful tool even when the text is unfamiliar or complex. Through repeated practice with multiple text structures and understanding the process of creating strategic memories, the learners may be able to impose structure to read and understand texts that lack cohesion and are dense. A web-based tutor was tapped to deliver consistent and high-quality instruction to all students.

The text structure strategy instruction begins with identifying signaling or linking words, classifying the text structures, summarizing with text structure-based scaffolds, encoding strategic memory structures, inferring, elaborating, applying, and writing. Specifically, the text structure approach used in this project is referred to as the *text structure strategy* and was developed by Meyer (1975) and systematically refined through multiple studies (e.g., Meyer et al., 1980; Meyer & Poon, 2001; Meyer & Wijekumar, 2007; Meyer et al., 2010). The delivery of the text structure strategy in this project used a web-based intelligent tutoring system designed to increase the likelihood that 7th graders received modeling, practice tasks, assessment, and immediate and scaffolded feedback to improve their content area reading comprehension mitigating any teacher factors (Meyer & Wijekumar, 2007; Wheldall, 2005). ITSS uses texts from science, social studies, current events, and sports so that disengaged learners may find some topics of interest (Guthrie & Davis, 2003) to practice the text structure strategy. Once students have become proficient in using the text structure to read, select, encode, and comprehend content area texts that are well-signaled they can extend the skills to real-life texts that may be poorly signaled, dense, and/or lack cohesion. ITSS uses well-signaled passages initially and transitions students to complex real-life texts to show students how to transfer their knowledge about the structure strategy to poorly signaled texts.

Text Structure-Based Reading Comprehension—Theory, Research, Practice, and Policy

Text structure-based instruction for comprehension has theoretical, empirical, and policy support. The theoretical basis for the text structure strategy grew out of research on linguistics, cognitive science, and educational psychology where a hierarchy of subordination of some ideas to others and discourse markers within expository texts were linked to memory representations of the text and improved outcomes on reading comprehension measures (Meyer et al., 1980). Learning this strategy enables readers to strategically build mental representations similar in organization to the author's organization (Gernsbacher, 1996; Meyer, 1975) or centrality of connections (Goldman, Varma, & Cote, 1996) to a text's hierarchical organization of important ideas (Meyer et al., 1980). The text structure strategy can be particularly helpful in unfamiliar domains of learning (Meyer, 1984; Voss & Silfies, 1996) and helps learners to begin building their knowledge base, a critical factor in learning to read content material (Alexander, 2005). Consistent with Alexander's developmental model of learning in academic domains, the text structure strategy can be a useful tool when students are in the acclimation state of learning and have to acquire knowledge about one or more domains through reading.

The foundations of the text structure model of comprehension proposed by Meyer (1975) share many elements with the construction-integration (van Dijk & Kintsch, 1983) and landscape models (Taylor, Graves, & van den Broek, 2000; Yeari & van den Broek, 2011). The shared foundations include the top–down processing, integration with prior knowledge, focus on memory structures, and interactions among reader, text, and task (e.g., Bohn-Gettler & Kendeou, 2014; Meyer, 2003; Meyer & Rice, 1989). The text structure-based strategic memory structure can been considered as an example of a type of situation model in the construction-integration model, where prior knowledge and new information are integrated (Meyer & Poon, 2001; Stine-Morrow, Gagne, Morrow, & DeWall, 2004). The variations of the approaches are mostly related to the implementation of these models during instruction about reading comprehension. For example, the implementations of the construction-integration model focus on summarizing, cohesion of text, and inferences; instruction often focuses on reading and rereading the text for summarizing with feedback given to scaffold the construction of effective summaries (e.g., Caccamise et al., 2014) and self-explanations (e.g., McNamara, 2004). The text structure-based approach is more explicit, precise, and transparent in scaffolding the reader's attention to the most important elements of the text through the main idea patterns for each text structure. Instead of repeated efforts to read the text, children receive specific instruction to look for who was being compared with whom and on what basis they were compared if the passage compared two or more people (e.g., Comparison pattern: _____ and _____ were compared on ____, ____, and ____). The construction-integration model focuses on activation and repetition of words and cohesion (Halliday & Hasan, 1976) at the sentence level or paragraph level. In contrast, the text structure model relies on the five text structures and hierarchical nesting of these structures to provide cohesion among the ideas within the text. Again, both the construction-integration and text structure models encourage inferences, elaborations, and effective utilization of memory, but the nature of instruction varies. Specifically, the text structure strategy provides scaffolding for the learner to infer and elaborate based on text structures. The strategy supports construction and integration and explicates the contents of a coherent situation model allowing direct and indirect scaffolding of reading comprehension.

The text structure-based model is also similar to the automatic and strategic cognitive processes underlying text understanding described in the landscape model (Yeari & van den Broek, 2011).

In addition to bottom–up features, Yeari and van den Broek (2011) stated,

> Landscape Model considers the organization of a discourse structure and its constituent segments (such as sentences, clauses, and propositions) as they guide the comprehension processes, and the role of linguistic cues (including connectives such as "therefore," "because," "after," "next to") as they direct the reader to maintain particular types of standards of coherence (Sanders, 1997). (p. 638)

Similarly for the text structure model the five basic text structures and hierarchical nesting of these structures guide the reader to select and encode information.

Text structure-based reading comprehension has been recommended by experts (e.g., Mallette, Duke, Strachan, Waldron, & Watanabe, 2013; Pearson & Hiebert, 2015), implemented in National and state policies, and included in textbooks (e.g., Scott Foresman Reading Street) as well as components of recent interventions (Vaughn, 2015). Additionally, text structure-based reading comprehension skills are reflected in the new grade-level expectations for reading in the Common Core State Standards Initiative (CCSSI; Gewertz, 2011; e.g., Reading CCSS.ELA-Literacy.RI.5.1 to RI.5.10) adopted by 42 states. Similarly, state standards (e.g., the Texas Essential Knowledge and Skills) also recommend the use of text structure at upper elementary and middle grades. However, there are important differences between the text structure strategy developed by Meyer and colleagues and these textbook and intervention approaches.

There are notable differences between the text structure strategy and most textbook uses of text structure (e.g., Foresman, 2007) or classroom interventions for reading comprehension, including those that add text structure to the instruction (e.g., Promoting Acceleration of Comprehension and Content Through Text, PACT; Vaughn, 2015). The text structure strategy subsumes other reading comprehension approaches under the text structure umbrella, provides explicit, precise, and transparent scaffolding, and is efficient. Explained another way, the text structure strategy instruction for reading comprehension may share some constructs with other comprehension curricula but the organization and roles of the instructional strategies and activities are different.

Interventions with a similar foci to the text structure strategy and designed to improve reading comprehension with middle grade learners include SERT (McNamara, 2004) and iSTART (McNamara, O'Reilly, Best, & Ozuru, 2006), Summary Street (Wade-Stein & Kintsch, 2004), and Computer-Assisted Collaborative Strategic Reading (CACSR; Kim et al., 2006), Learning Strategies Curriculum (LSC; Cantrell, Almasi, Carter, Rintamaa, & Madden, 2011) and PACT (Vaughn, 2015). Summary Street, SERT, and iSTART are supplemental classroom interventions that trace their roots to the construction-integration model and focus on summarizing and self-explanations as the means to achieving deep comprehension. A series of interventions designed for struggling adolescent readers also present limited text structures as part of their curriculum (i.e., CACSR, LSC, and PACT).

The differences between the text structure strategy and these interventions and curricula are listed below.

1. Overarching organization of language arts instruction using the five text structures as a guide to manage the selection of important ideas, encoding of strategic memory, and utilization of memory in writing summaries, generating inferences, extrapolating and extending knowledge structures, and writing. This allows the relationships between the ideas to become the organization structure and hierarchical and efficient memory guide. Other implementations of text structure present tasks such as summarizing separate from the text structure instruction.

2. Scaffolded summary writing tasks based on the text structure. This is particularly useful to novice learners. The scaffolds can be gradually released when students become proficient.

3. Scaffolded inference, elaboration, and comprehension monitoring based on the text structure-based relationships between the ideas.

Description of the Text Structure Strategy Intervention and Web-Based Delivery

The ITSS focuses on cognitive and metacognitive skills necessary to support the selection and encoding of important elements of the text into a coherent mental representation. The ITSS guides these activities as outlined below.

1. Identifying the organization of the text as one of five text structures (individually or nested). The reader can use the authors' intended text structure if it is signaled or impose structure when no signals are present (e.g., authentic texts that are poorly signaled).

2. Scaffolding the selection of the most important elements in the text to write a main idea. In the problem and solution text structure the goal is to highlight the problem(s) and solution(s). The problem and solution main idea scaffold is: "The problem is ______ and the solution is ______________." Students can add as many blanks as needed to extend the main idea. For example, the passage about the problem with garbage shown in Figure 2 was used in the teacher professional development in the recent trials. The article was adapted from "Howthingswork.com" and provides a problem and multiple solutions to the problem. The main idea for this article using the text structure strategy is "The problem is *garbage* and the solutions are *recycling, using less resources, and incineration."*

3. Promoting the creation of a strategic cognitive structure for the passage using the main idea. If the student has some prior knowledge then their prior knowledge can be revised and updated to include the new information. Figure 3 presents one example of the reader's strategic memory representation developed through the main idea scaffolding/pattern.

4. Supporting comprehension monitoring and checking memory structures using the text structure's organization (e.g., Do I remember the solutions to the problem?)

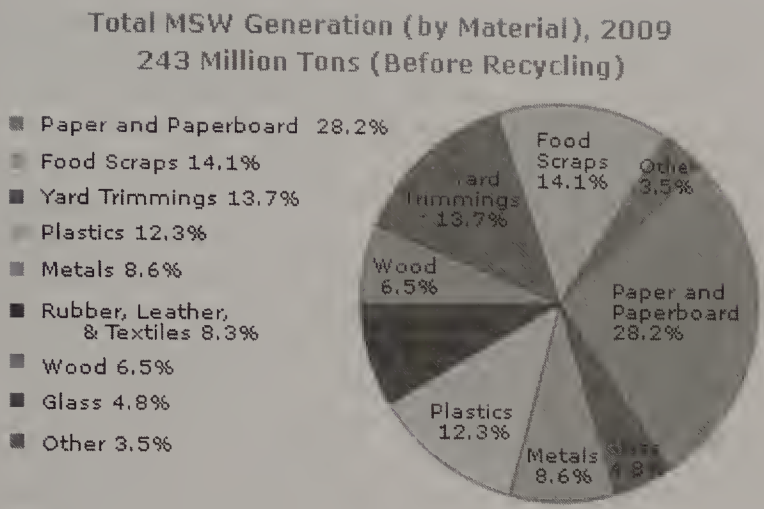


Figure 2. Garbage lesson text excerpt. From Municipal Solid Waste in the United States: 2009 Facts and Figures, US EPA, Washington, DC. *The Trouble With Trash* (Text adapted from: http://www.howstuffworks.com/landfill.htm): "Americans generate trash at an astonishing rate of 4.6 pounds per day per person, which translates to 251 million tons per year! What happens to this trash? Some gets recycled or recovered and some is burned, but the majority is buried in landfills. Of the 251 million tons of trash, or **solid waste**, generated in the United States in 2006, about 81.8 million tons, or 32.5 percent, was either recycled (glass, paper product, plastic, metals) or composted (yard waste), and 12.5 percent is burned. The remaining 55 percent is discarded in landfills and made up of mostly paper products, food scraps, yard trimmings, & plastics. The pie chart below shows the percentage by which different materials contribute to the municipal solid waste stream." Freudenrich (2000). The amount of trash buried in landfills has doubled since 1960. This is a problem, because landfills are not designed to break down trash, merely to bury it. Trash put in a landfill will stay there for a very long time. Inside a landfill, there is little oxygen and little moisture. Under these conditions, trash does not break down very rapidly. In fact, when old landfills have been excavated or sampled, 40-year-old newspapers have been found with easily readable print. When a landfill closes, the site, especially the groundwater, must be monitored and maintained for up to 30 years! In many areas worldwide, landfill space is running out. This is due to people not wanting landfills near their homes. If changes aren't made, a landfill shortage crisis will happen within the next 10 years. There is a way for us, as consumers, to help out in the fight against pollution. One solution is that we can practice the three R's: Reduce, Reuse, and Recycle. Sometimes people throw out items that are expensive to get rid of that can be reused. For example, the cost of disposing one barrel of oil-based paint is $630–$1,200 and the paint could be used by someone else. If people want to stop the landfill crisis before it begins, they should work harder to reduce garbage, donate items that can be reused by someone else, and recycle more.

This approach promotes a top–down process for reading comprehension that is strategic in processing, efficient due to chunking, and well-associated through the five text structure patterns and nested structures. For example, as the readers develop their *memory structures* for the *problem(s)* and *solution(s),* they can *extend* their memory to associate the *cause(s)* for the problem(s) thereby creating more linkages that prove to be well-associated. The learner can also be prompted to *infer* the causes to the problem with the question: "Can you figure out possible causes for the problems based on what you know about garbage or what you can research?" At this point the memory structure (see Figure 4) can be extended to include the causes for the problem. Figure 5 shows an elaboration of memory structures in Figure 4 where a comparison

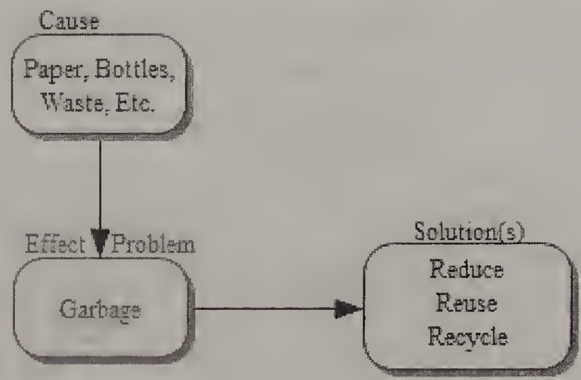


Figure 3. Possible strategic memory structure for the passage.

of solutions is nested within the problem and solution overarching structure. ITSS lessons included passages showcasing nesting of two, three, and four text structures.

The text structure strategy serves as a metacognitive approach prior to (e.g., planning to read), during (e.g., extraction of main ideas), and after reading (e.g., comprehension monitoring) as presented in a series of YouTube videos designed and developed by Wijekumar (2014a, 2014b). Prior to reading, the reader can plan to use text structures to impose top–down structure on their reading whether the passage is signaled or not. For example, after skimming the text or reading the heading, the reader can decide to use the problem and solution text structure. The reader can then approach the text strategically seeking information about the problem and solution and use that same approach to bring cohesion to the text. During comprehension, the text structure strategy's main idea patterns guide the selection and encoding of information in meaningful and associated chunks. Either during or after reading, children can monitor their comprehension by traversing the main idea based memory structures to confirm synthesis and understanding of the text. When reading about a problem, the reader can reflect back on the passage they read and check whether they remember the problem, solution(s), and cause(s) for the problem. This approach also allows students to detect and repair any inconsistencies or fill in missing information. These monitoring and

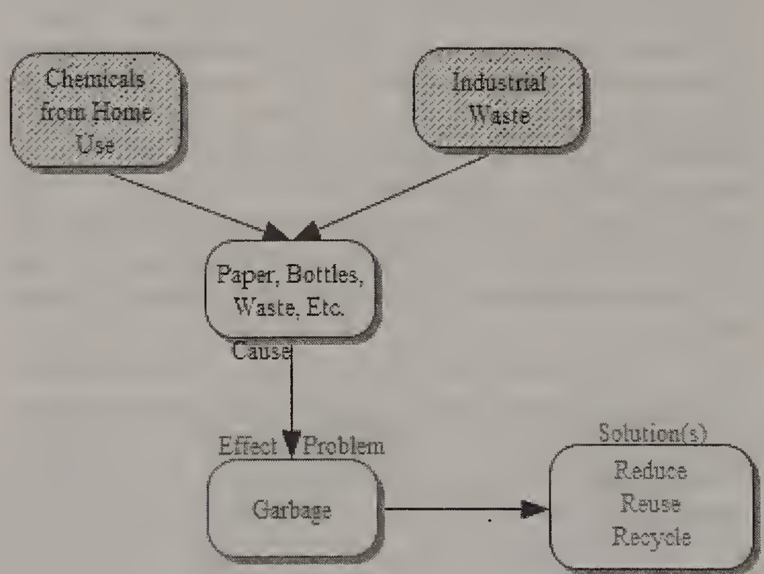


Figure 4. Possible inferences about causes for the problem and enhanced memory structure about garbage.

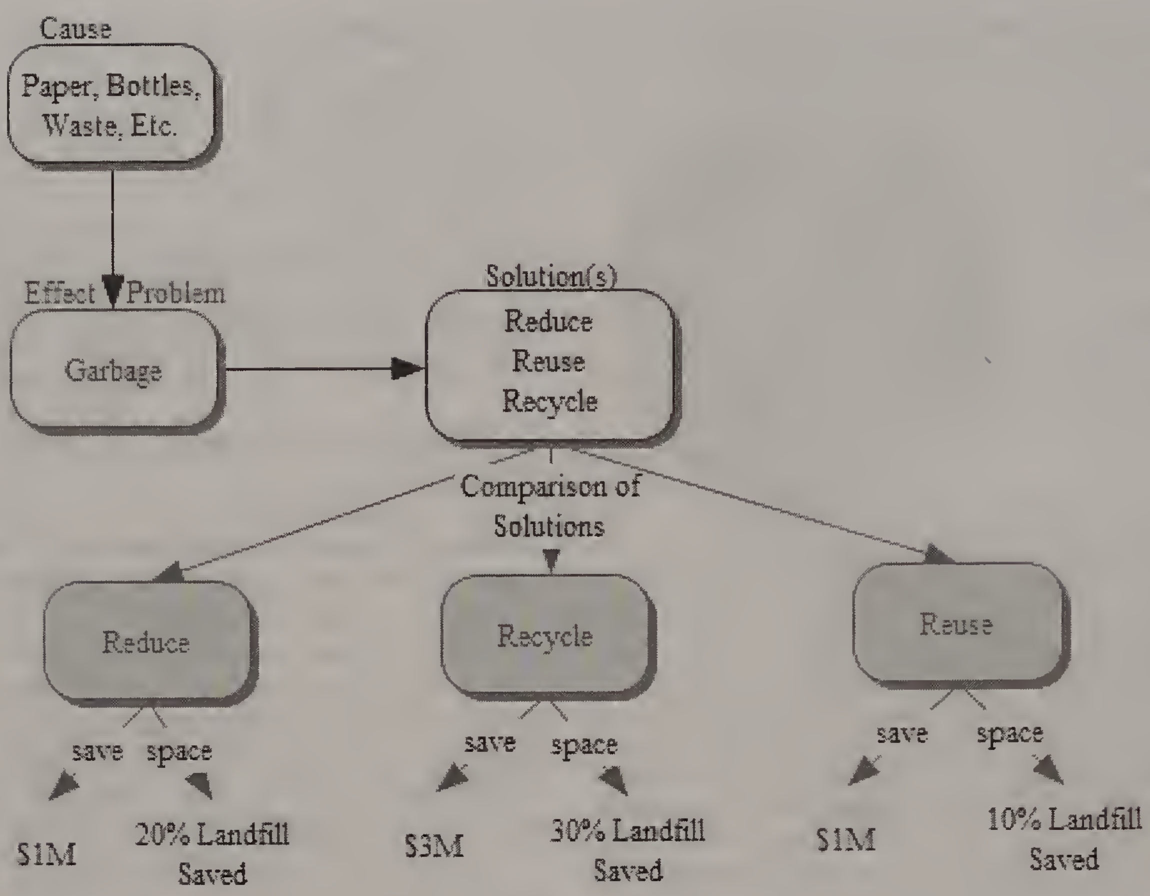


Figure 5. Nested text structure memory—comparing solutions to the problem (estimates for costs).

revising/repairing activities may require further information seeking, inferences, or elaboration and can be scaffolded by the text structure.

The web-based delivery of the text structure strategy in ITSS supports learning through modeling of strategy use, practice tasks (e.g., writing a main idea using the pattern for a particular text structure, identifying texts that inform or persuade), assessment, and scaffolding feedback with different types of expository texts from different content domains. Immediate feedback with hints and further modeling of good levels of understanding show students how to set and maintain standards of coherence for understanding expository text in classroom learning settings. Thus, the text structure strategy serves to establish standards of coherence with learners who are most likely to overlook them due to a lack of prior knowledge, limited working memory, inability to maintain standards of coherence, and/or difficulty taking advantage of important linguistic cues in the text.

ITSS meets 11 of the 15 recommendations by Biancarosa and Snow (2006) (e.g., direct, explicit comprehension instruction, strategic tutoring, diverse texts, and technology component). An animated pedagogical agent, named I.T., initiates instruction, guides the learner with information on practice tasks, presents feedback, and supports the learner through multiple attempts at learning the text structure strategy. The ITSS platform also maintains extensive records of student progress allowing teachers to view reports and manage the classroom delivery of the approach. Typically, a student logs into the system and receives instructions from I.T. The student proceeds at his or her own pace by listening to and viewing I.T.'s model, reading texts on the screen, answering questions (with multiple trials), and getting hints and feedback from I.T. The narration may be turned off for higher grade level students if the teacher believes it is not necessary. The students read passages from science, social studies, sports, and current events of varying lengths and reading levels. Learners first receive instruction about the comparison text structure followed by the problem and solution and cause and effect text structures. They also receive instruction on combining or nesting text structures. Passages used in ITSS include those created by Meyer and authentic texts from real-life sources, such as newspapers, magazines, and online sites. There are over 100 lessons within ITSS and many versions of the lessons with easier passages for readers experiencing difficulty reading. ITSS provides direct and explicit comprehension instruction, diverse texts, well-tested technologies, ongoing formative and summative assessments, and 2–4 hours of professional development for teachers.

ITSS is designed to work as a partial substitute to the language arts curriculum and is typically delivered once a week for approximately 30–45 min. Ideally, the system is used in concert with the teacher-managed reading comprehension instruction that takes place on days when the computer software is not used. The text structure strategy instruction delivered via the web-based ITSS has empirical support at the elementary-grade levels (e.g., Wijekumar et al., 2014, 2012). In a pretest posttest design with equivalent forms of tests counterbalanced across testing times (Meyer et al., 2010), 56 fifth and 55 seventh graders substantially increased their standardized reading comprehension test scores, use of the strategy, amount of information remembered from reading, and identification and recall of main ideas. For example, after 6 months (90 min/week) of this ITSS instruction with advanced feedback the average improvement for students reading below grade level was

2 grade equivalent levels for the 5th graders and 3.8 grade equivalent levels for the 7th graders (Meyer et al., 2010). Lower-achieving readers in 7th grade showed the highest level of improvement from pretest to posttest (d = 1.85). The below grade-level readers in 7th grade doubled their recall after ITSS instruction.

ITSS has not been tested in a large-scale study at the middle grades but has been efficacious at 4th and 5th grade as evidenced by findings from two recently completed multisite cluster randomized controlled trials with approximately 260 classrooms at Grades 4 and 5 (Wijekumar et al., 2012, 2014). Results showed statistically significant and meaningful impacts on 4th and 5th-grade students reading comprehension on both the standardized Gray Silent Reading Test (GSRT; Wiederholt & Blalock, 2000) and researcher-designed measures. Fifth graders in ITSS classrooms on average scored 0.2 SDs ($p < .05$) higher on GSRT adjusted posttest scores and .42 SDs ($p < .05$) higher on comparison signaling posttest scores than 5th graders in control classrooms holding reading pretest scores, gender, and school locale constant. Also adjusted posttest scores were statistically significantly higher for 5th-grade students in ITSS classrooms than their control counterparts on all other researcher measures: main idea quality Effect Size (ES = 0.53), comparison total recall (ES = 0.32), and comparison competence (ES = 0.26; Wijekumar et al., 2014). When 4th- and 5th-grade teachers implemented the text structures consistent with the ITSS approach, the results showed larger effects (Wijekumar, Meyer, & Lei, 2013).

ITSS yielded more benefit on the standardized test of reading comprehension for below grade level 4th-grade readers with the greatest needs for improvement in reading comprehension. Due to teachers' concern about typing skills, the ITSS for 4th grade was truncated to focus on using the text structure strategy to construct strong main idea statements (about two sentences) after reading texts, rather than constructing both main ideas and recalls in the complete version of ITSS used with 5th graders and above. The main ideas in ITSS can be seen as situation models organized with text top-level structures (e.g., problem and solution) and focusing on macropropositions, rather than micropropostions within a sentence or between adjacent sentences. The Wijekumar et al. (2012) study provided support that 4th graders reading below-grade level can learn to write good main ideas using text structure to integrate important ideas across two paragraphs of expository text. Additionally this learning transfers to reading comprehension performance on a standardized test.

In studying students without ITSS instruction in a cross-sectional design across grades four through nine, Meyer, Ray, and Middlemiss (2012) found that the largest growth across grades in understanding the comparison text structure was for average comprehenders and no improvements past 6th grade were found for low comprehenders. In the current study, we examined whether the ITSS intervention can increase understanding and use of the comparison text structure by 7th graders, including low comprehenders.

Finally, in previous studies of ITSS exploratory analyses were conducted to study the effects of time on task and numbers of questions answered by the participants. At the fifth grade level students who answered more questions showed better performance on the outcome measures but average minutes used was not significantly related to GSRT posttest scores (Wijekumar et al., 2014).

The Current Study

Consistent and high quality instruction in reading comprehension may be difficult to achieve in large numbers of classrooms due to location and other challenges. Students attending rural or suburban schools may have access to different resources and may be socioeconomically different. Schools may also differ based on curricula, teacher quality, and student background. The web-based ITSS was designed to overcome these challenges and provide consistent modeling, practice tasks, assessment, scaffolding, and feedback to the learners.

In this large-scale randomized controlled efficacy study, we examined the effects on reading comprehension of 7th-grade students learning how to select, encode, and strategically organize text via the ITSS versus traditional 7th-grade reading comprehension instruction that focuses on activities that emphasize summarization, questioning, and highlighting of texts independent from the text structure framework. These activities were identified during a review of the curricula used in the schools. Based on the theory and supporting research studies, we hypothesized that 7th-grade students learning to read and comprehend expository texts using the ITSS will outperform their business as usual counterparts who do not use the ITSS.

We also conducted exploratory analysis of five factors that may affect the intervention's outcomes variably, such as initial skills (Stanovich, 1986), gender (Halpern, 2006), and school and time factors. Interventions often show larger effects for more skilled readers, where the rich become richer as described in the Matthew effect (Stanovich). However, as noted earlier, Wijekumar et al. (2012) showed the opposite effect with below-grade level 4th-grade readers benefiting more from ITSS than more proficient readers. Low comprehenders by 7th grade may have experienced no growth in understanding how ideas can be related via text structures among sentences and paragraphs. Additionally, years of failure constructing useful memory representations after reading may have left them defeated with negative attitudes toward reading to learn. It is an important question to see whether different types of comprehenders benefit differentially from ITSS at the 7th-grade level. Concerning gender, Halpern (2006) explained stronger performances of females in comparison to males on some reading comprehension and writing tasks using complex prose as well as retrieval from long-term memory and writing lengthy responses. Finally, the amount of time spent on the web-based software and the numbers of questions completed by students have been shown to matter in previous studies. Therefore, we explored time on task and number of questions completed in this study as well.

Research Questions

This study was designed to answer the following primary research question. Do students in Grade 7 classrooms using the ITSS delivery of the text structure strategy as a partial substitute for the standard language arts curriculum outperform students in control classrooms on standardized and researcher-designed measures of reading comprehension?

The study also posed five secondary questions concerning whether the effect of ITSS delivered instruction about the text

structure strategy for reading comprehension varies depending on other factors, including reading skills, gender, schools, setting (i.e., rural vs. suburban), and time working in ITSS. The five secondary questions are as follows:

1. Does the effect of ITSS on reading comprehension depend on students' initial reading level?

2. Does the effect of ITSS on reading comprehension differ between male and female students?

3. Does the effect of ITSS on reading comprehension vary across rural versus suburban areas?

4. Does the effect of ITSS on reading comprehension vary across schools?

5. Do students who used the ITSS system more in terms of time or number of answered questions perform better on the posttest than students who used it less?

Method

Design

This multisite cluster randomized efficacy study investigated the effects of a web-based tutoring system (ITSS) to deliver the structure strategy-based comprehension instruction to 7th-grade students in rural and suburban settings. A volunteer sample of 108 classrooms were stratified by school and randomly assigned to ITSS or business as usual control. Schools agreed to use the ITSS software as a partial substitute for the language arts period for 30–45 min each week. During the intervention time, the ITSS classrooms used the web-based tutor delivered instruction with modeling, practice, assessment, and feedback for each student individually. The ITSS classroom teachers delivered the school's language arts curriculum for the rest of the language arts instructional time. The within-school random assignment of classrooms meant that the business as usual control teachers used the school's standard language arts curriculum for the total language arts instructional time.

The within-school random assignment of classrooms required fewer participating schools (compared with random assignment of schools) and provided curricular consistency between the ITSS and control classrooms. Schools were also eager to participate because they would build capacity to use the intervention, and control classrooms had the opportunity to use the software after the posttests were completed. The possibility of contamination was minimized by the password protected ITSS software for the students.

Participants

A team of laboratory-extension specialists from a large research university in the Northeast led by the project director recruited schools to participate in the study by sending letters of invitation to all schools within two states and following up with phone calls and regional presentations to school leaders. Requirements to participate in the study included the availability of computer labs with high bandwidth network connections to the Internet. All schools in both states met the requirements because of recent one-to-one computer initiatives and wide area networks. The research team completed site visits to all interested schools and verified the availability of the computers and bandwidth and received approval from the school administrators.

The recruitment effort resulted in a total of 25 schools (14 rural, 11 suburban) agreeing to participate in the research study. The recruitment was completed in two cohorts. These schools had an average student to teacher ratio of 14:1 in both rural and suburban settings based on school districts data reported on state websites. The average class size in the participating classrooms was 21. The average educational expenditure rate was $13,874 per student. The schools' student population was 8% racial/ethnic minorities and 42% socioeconomically disadvantaged (eligible for free and/or reduced priced lunch).

Incentives to participate in the study included the professional development for teachers and the free use of the ITSS software for the study year as well as a second year. Teacher aides were recruited by the schools and paid by the research funds to assist in the setup of the computer labs and monitor usage during the intervention delivery.

All 7th-grade language arts teachers in the participating schools were invited to participate and none declined. Middle school language arts classes were organized with one teacher teaching multiple classes. The random assignment was done at the teacher level. As a result we had 59 classrooms in the ITSS group and 49 in the control group after random assignment of the teachers within each school. All the participating teachers (classrooms) were randomly assigned to the ITSS (intervention) or control conditions after students had been assigned to teachers in the schools. Teachers completed their consent forms at the professional development sessions or during the site visits by the study team. All students in the 7th grade of participating schools were invited to participate. Each school mailed parental consent forms to all students at the 7th-grade level prior to notification of random assignment. Student consent was obtained at the pretest sessions and 96% of students agreed to participate.

The analysis sample consisted of 2,489 7th-grade students from a total of 108 classrooms from 25 schools. Students who used the software for a total of 30 min or less throughout the year were excluded from the analysis. Many of these students were receiving pull-out special education services during the ITSS time. The determination was made by the schools. The special education students in the control classrooms also received pull-out instruction and did not participate in the study. About 48% of the student sample (48.2%; n = 1,200) was female, 56.9% (n = 1,415) were in the ITSS condition, and 53.2% (n = 1,325) came from rural districts.

Procedure

Measures of reading comprehension (standardized reading comprehension test followed by researcher-designed measures) were administered (to both ITSS intervention and control groups) during the pretest before training began. The testing sessions were conducted by members of the research team in the presence of the teachers in the school auditorium or cafeteria. Teacher professional development was delivered by the research team to the intervention teachers at the beginning of the academic year. The session

lasted approximately 3 hours and provided the teachers with a description of the text structure strategy, showed how ITSS functioned, and described typical student interactions with the software.

Schools agreed to allow the students to use the ITSS software for one or two sessions a week for 30–45 min each week over a 6- to 7-month period during the school year as a substitute for the regular language arts curriculum. Teacher aides were hired by the research team to ensure the smooth implementation at each school. The teacher aides were present during the computer lab time and notified the research team of any computer, bandwidth, or implementation issues. At the beginning of each session, each student picked up their ITSS folder containing any instructions, username, password, and earphones and sat individually at the computer. The student opened a browser and logged in using their individual username and password. The ITSS software initiated the interactions with the student by starting the new session based on the last completed lesson and activity. Students interacted with the ITSS program at their own pace, listening to I.T., responding to questions (e.g., click on signaling words, write a main idea), and receiving feedback and help from I.T. At the end of the class period students logged out and their work was saved.

The ITSS instruction focused on how to (a) identify the text structure(s) (b) select and encoding information strategically when reading, (c) use the top-level structure to write a good main idea, and (d) use the five text structures and nested structures to remember the important text information and details. When students responded to questions the system assessed their response and selected appropriate responses from the database based on the score, attempt/try number, and type of question (e.g., on the second attempt for a main idea question type the learner may receive audio only feedback that says, "You have only written who was being compared but need to add information about what they were compared on." If students move to a third or fourth attempt at the same question, they may see a pop-up window showing them more information that they should include when they revise their main idea.) The assessment system reviewed responses for nonsense words, blank answers, repeating the same answer, and words from the urban dictionary (using an application program interface) to detect gaming and notified the teacher about the activities by flagging those words in the reports.

Fidelity of the treatment was monitored by the research team through classroom observations and weekly review of ITSS computer logs. One classroom observation was conducted during the year in both intervention and control classrooms and noted the types of instructional activities, overall classroom atmosphere, and classroom organization (e.g., small group, teacher-led). The observations documented any use of text structure and other comprehension strategies in both the intervention and control classrooms and noted similarities and differences between the ITSS version of text structure versus other approaches (as described earlier). The observations also noted any possible contamination of the control classrooms.

Biweekly progress reports were emailed to the teachers in the intervention group noting student progress and any gaming of the system by students. If students submitted nonsense or blank answers repeatedly or used language in the urban dictionary, the system flagged the interactions as gaming and teachers were asked by the research team to follow-up with the student(s). Alerting students that teachers would see their written responses along with teacher follow-up reduced off-task behaviors in the ITSS (Wijekumar et al., 2014).

Posttest measures on reading comprehension were administered at the end of the school year under the same conditions as the pretest. Posttests included the GSRT and researcher-designed measures. When students had to leave early from any testing session, the research team advised them to complete the standardized test and the signaling word task of the researcher designed measure prior to leaving.

Materials

Materials for this project included the web-based lessons described earlier and teacher professional development materials (i.e., PowerPoint description of text structure, video on how ITSS functions, and sample lessons). The measures administered at pretest and posttest are described below.

Reading comprehension outcome measures. Standardized and researcher designed measures for cognitive outcomes were administered at pre and posttest.

Standardized test of reading comprehension. The GSRT (Wiederholt & Blalock, 2000) was used as the standardized distal measure of reading comprehension. There are two forms of the measure, forms A and B, and each uses 13 progressively longer and more difficult narrative texts with five multiple choice questions for each passage. The questions range from passage independent questions that rely on prior knowledge, locating information in passage, elaborative, cohesive, and knowledge-based inferences, and vocabulary dependent types. The ProEd (2015) website notes that "reliability Coefficients Alpha are all at or above .97." We also studied test-retest, alternate forms-immediate, alternate forms-delayed, and scorer reliability. Cronbach's alpha for both forms of the GSRT was reasonably high ($\alpha = .88$). During this study the GSRT Form B was administered at pretest and Form A was administered at posttest. The pretest GSRT score was used as a covariate for data analyses when examining the effects of ITSS instruction on our dependent measures focusing on reading comprehension. The posttest GSRT score was the outcome for the primary research question.

Researcher-designed measures of reading comprehension. Two equivalent test forms designed to measure student use of problem and solution and comparison text structures were created (Meyer et al., 2010). One form was administered before the students started ITSS and the second immediately after completing the program. Each form had three passages: one using the problem and solution text structure, one short comparison text structure and one long comparison text structure passage. The problem and solution and short comparison texts were used in the randomized controlled trials conducted with fourth and fifth graders (Wijekumar et al., 2012, 2014). *Top-level structure* and *competence* were gathered for both the problem and solution and comparison passages. *Signaling word* identification was measured using the short comparison passage. Both short and long comparison passages have an additional variable on *number of issues compared*. The passages and measures are described next.

The comparison and problem and solution text structures were selected for measurement in this and previous studies. Both those

text structures provide a rich platform for showcasing the power of text structure in selection and encoding of hierarchical memory structures. They are less frequently used in classroom instruction than the sequence and description structures that are less efficient in organization with fewer opportunities for being strategic and chunking (Meyer & Freedle, 1984). Because this research was an efficacy study designed to test the ITSS system under optimal implementation conditions the comparison and problem and solution text structures matched the sequence of lessons within the ITSS system where 12 comparison text structure lessons were followed by 10 problem and solution lessons and another two review or extension lessons with both text structures. In an efficacy study the proximal measures should be closely aligned to the instruction, and as such, we also anticipated that it is most likely that students in the ITSS condition would have encountered instruction in these two text structures prior to the posttests.

Problem and solution text structure passage. Two passages for the problem and solution structure were prepared: (a) rats (authentic newspaper article, see Meyer & Poon, 2001) and (b) dogs. The two equivalent passages had 98 words, 72 idea units, and equivalent scores on traditional measures of readability, text structure, and signaling (see Meyer, 2003). Each text presented a relatively unfamiliar problem and its cause and a solution that eliminated the cause of the problem. Students were asked to recall all they could remember after reading each problem and solution text and placing it out of sight in an envelope. Dependent variables for the problem and solution texts included the *top-level structure* and *competency* of using the problem and solution to organize the recall.

Short-comparison passages (CO-Short). Two short passages were also prepared for the comparison structure: (a) pygmy versus Emperor monkeys and (b) Adelie versus Emperor Penguins. Each comparison passage had 128 words, 15 sentences, and 96 idea units. There were three tasks for the comparison structure: (a) a fill-in-the-blanks cloze task to complete 4 blanks in the short comparison passage, called the signaling test, (b) a recall task like that used for the problem and solution set of articles, and (c) a comparison main idea task where the student was asked to write a two-sentence main idea with the text available for consultation. Dependent variables for the short-comparison texts included *top-level structure* and *competence* similar to the problem and solution set of texts, and *number of issues compared* and *signaling test* scores.

Long comparison passages (CO-Long). Two longer comparison text structure passages were also created and used at pretest and posttest, respectively: (a) Hagar Qim Stone Circles versus Stonehenge (text about Hagar Qim and Stonehenge adapted from Hammann, 2000), and (b) Mt. Rushmore versus Easter Island. Each comparison passage had 527 words, 33 sentences, and 134 idea units. The same scales and procedures were used as for the short-comparison texts for recall: *top-level structure*, comparison *competency*, and *number of issues compared*.

Scoring

Scoring was done using computer algorithms for the signaling word responses and trained raters for the top-level structure, competence, quality, and number of issues compared measures. The short-comparison fill-in-the-blanks answers were scored by a computer algorithm and correct answers were given a score of 7 for each response with a maximum possible score of 28.

Comparison and problem and solution *competence* from the main idea and full recall tasks were scored by two trained raters supervised by a skilled researcher using manuals developed for two previous research projects (Meyer et al., 2010; Wijekumar et al., 2014). Competency ratings for use of the problem-and-solution and comparison structures (proximal measure with scores from 1 to 8) were assessed to determine the degree to which a 7th-grade student proficiently used the text structure as outlined in the ITSS program (i.e., correct problem in the text with cause and its correct solution). These scores were based on the full recall of the text without the passage in view. For recalls from the comparison texts students presenting both elements compared, issues contrasted, and correct details of several of the issues contrasted received a score of 8. Students presenting some details without any organization received a score of 1. The same scoring was used for the comparison main idea task except that a 6-point competence scale was employed rather than 8-points scale. The short main idea required only two issues for the highest competence or quality score of 6; one issue could use words from the text insight during writing the main idea, such as "feed on fruits," but a second issue required using a semantically superordinate category generated by the adolescent, such as "diet of fruits." Any correct issue compared for the two elements/creatures received a score of 5. A complete breakdown of the different scores with examples are presented in Wijekumar et al. (2013). Scoring was based on a propositional analysis of the ideas in text with interrelationships among ideas specified in a hierarchical content structure. At least 10% of the data from each of the measures were randomly selected from the conditions and time of testing to check interrater agreements. All scorers were blind to treatment conditions as well as factors of secondary interest in the study. Intensive training and mentoring were provided for pairs of educational psychology graduate students, who separately scored each protocol until they could independently score with at least 90% agreement. Then scorers were randomly assigned protocols to score, which included a randomly selected 10% of overlapping protocols for continuous weekly reliability checks. Weekly scoring checks for pairs of students were mentored and monitored by an experienced faculty researcher to prevent drifts in scoring and ensure high consistency and reliability in scoring. Most scorers had two to three years of experience with the research team. The longer comparison texts for 7th-grade students was new to the scoring team and was scored by the experienced faculty researcher and a school psychology graduate student; training was intensive and reliable. For example, the final 10% check for the posttest showed agreement between the scores of 95.30%, 95%, and 97.30%, respectively for top-level structure, comparison competence, and number of issues compared.

The percentages of agreements between scorers for competency scores ranged from 86.3% to 95.8%. Agreement for comparison main idea competence ranged from 96.9% to 99.5%. Percentage agreement for the number of issues compared ranged from 96.7% to 100%.

Recalls of problem and solution texts were scored for *top-level structure* (correspondence between the organization of the recall and the problem and solution organization of the text). For exam-

ple, a good top-level structure score (6 or higher on a 9-point scale) requires a problem part and a solution part (see Meyer et al., 2010, p. 80), but the solution does not have to be the same solution as that posited in the text. Scores greater than six for top-level structure include use of signaling words for the problem part (7 points), the solution part (8 points), and both the problem part and the solution part (9 points). At the low end of the top-level structure scale students only provide a descriptive list of ideas about the text with no indication in any of the sentences about the problem and solution structure (2 points). A score of 4 is also a descriptive list of ideas, but one of the listed descriptions shows the relationship between a problem and a solution.

Additionally for the short and long comparison texts, scorers tallied the number of issues correctly contrasted between the two objects (e.g., Emperor vs. Adelie penguins). There was high interrater reliability for the measures collected for this measure of number of issues compared (88%–100%). Two students with only slightly better than average performances on the pretest main idea task included a) 7th-grade student one: "the main idea is comparing the two monkeys and their differences," and b) 7th-grade student two: "Pygmy, and Emperor monkeys are different from each other." Seventh-grade student one's main idea was scored as a top-level structure of 4 out of 9 points possible, indicating some knowledge about the comparison structure, but not using the structure strategy to contrast two creatures on at least one issue. The competence was scored 3 out of 8 because the names of the two creatures compared were not identified (i.e., Pygmy monkeys vs. Emperor monkeys). The number of issues compared was scored 0. Similarly, 7th-grade student two's main idea received a top-level structure of 4, competence score of 4 for correctly identifying the creatures compared, and a main idea number of issues score of 0. On the posttest student one wrote, "Emperor penguins are larger than the Adelie penguin. They both live on Antarctica's pack ice"; this student's posttest scores were 6 for top-level structure, 5 for competence (two issues were worded similarly as those found in the text, but there was no generation of a superordinate issue), and 2 for issues compared (size indicated by larger and where they live). On the posttest 7th-grade student two wrote, "Emperor penguins are being compared with Adelie penguins by size, their growth, weight, appearance, diet, and where they live." This student received the maximum top-level structure and competence scores of 9 and 6, respectively. The number of issues compared were tallied for the main idea # of issues score; this student scored 6, one for each issue listed.

Data Analysis

Data analyses were conducted for each of the primary dependent variables (GSRT and researcher-designed measures of reading comprehension) using the HLM7 software program. Missing data was handled using listwise deletion at the time of analysis for each model to maximize the use of available data. Listwise deletion was used because missing at the classroom-level was relatively small (one class, <1%, missed reading pretests; 10 classes, <10% and 5 from each experimental condition, missed only researcher-designed posttests) and missing was not significantly associated with any of the observed variables at the class level. Furthermore, there was no statistically significant differential attrition between treatment and control conditions at the class or student level.

The amount of missing data at the student level was somewhat larger. About 6.3% of students (n = 156) did not participate in the GSRT pretest, 9.7% (n = 241) did not participate in the posttest, and 9% (n = 224) did not participate in either pretest or posttest. As noted earlier, the students were asked to complete at least the GSRT and the signaling word task of the researcher-designed measures prior to leaving the testing session. For the other researcher-designed measures, 152–154 students (6.1%–6.2%) missed just pretest, 377–382 (15.1%–15.3%) missed just posttest, and 239–240 (9.6%) missed both pretest and posttest scores (see Table 1 for the number of participants who completed each of the tests). Students missing reading posttest scores had slightly lower reading pretest scores, suggesting that missing might not be completely at random. However, reading pretest scores were included in all analysis models to mitigate the possible bias due to missing data (Graham, 2009). Students in the middle grades have schedules that did not align with the testing window and some had to leave early, and others could not be tested altogether. As the sample sizes for complete-case analysis remained fairly large at both the student and class levels, loss of statistical power was not a great concern.

HLM Model Specifications

A series of three-level hierarchical linear models (HLM; Raudenbush & Bryk, 2002), in which students were nested within classrooms within schools, were specified to address the primary and secondary research questions. An unconditional model (M0) was first estimated to gauge the outcome variability at each level. A main effect model was then estimated to answer the primary research question, in which there were predictor variables at each level. Student-level predictors included gender (1 = female, 0 = male; grand-mean-centered) and reading comprehension covariates. Reading comprehension covariates included group-mean-centered pretest scores on GSRT and a researcher-designed measure (i.e., signaling for the GSRT posttest outcome, or the corresponding pretest for researcher-designed outcome measures). Treatment efficacy was tested at the classroom level using contrast codes for experimental conditions (i.e., [1/2] = ITSS, -[1/2] = control; these contrast codes were used such that unstandardized regression coefficient corresponded to the difference between the unweighted means of the experimental groups). Classroom-level covariates included grand-mean-centered class average pretest scores on GSRT and the corresponding researcher-designed pretest measures. Differences between rural and suburban schools were examined (1 = rural, 0 = suburban; grand-mean-centered) at the school level. Variance associated with each of the three levels was estimated. This three-level main effect model (M1) was used to address the primary research question of whether 7th-Grade ITSS classrooms outperformed control classrooms on reading comprehension after controlling for other relevant factors such as prior reading level, gender, and school locale.

Each of the secondary research questions was addressed in a separate model by adding relevant interaction term(s) or random effects to M1. Specifically, cross-level interactions between treatment and each of the reading pretests (GSRT and corresponding researcher-designed pretest measure) were added to M1 by spec-

Table 1
Grade 7 Class- and Student-Level Means and Standard Deviations on Reading Measures

	ITSS						Control					
	Pretest			Posttest			Pretest			Posttest		
Measure	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Class level												
GSRT	59	35.75	5.02	59	39.71	4.74	48	35.18	5.48	49	36.82	6.38
Short comparison text												
Signaling test	59	13.82	2.69	59	15.77	3.28	49	12.87	3.74	49	13.08	4.43
Top-level structure	59	5.41	0.70	54	6.61	0.78	48	5.22	0.78	44	5.78	0.64
No. of issues	59	1.66	0.52	54	2.10	0.55	48	1.51	0.58	44	1.73	0.54
Competence	59	4.37	0.84	54	5.26	0.81	48	4.14	0.91	44	4.74	0.83
Long comparison text												
Top-level structure	59	2.67	0.67	54	4.65	0.92	48	2.67	0.58	44	4.07	0.79
No. of issues	59	0.20	0.19	54	0.68	0.37	48	.17	0.15	44	.45	0.28
Competence	59	1.48	0.47	54	2.77	0.66	48	1.44	0.36	44	2.38	0.61
Main idea												
Top-level structure	59	3.75	0.41	54	4.90	0.62	48	3.78	0.42	44	4.07	0.53
No. of issues	59	0.16	0.12	54	0.77	0.47	48	.19	0.13	44	0.22	0.18
Competence	59	2.97	0.42	54	3.84	0.44	48	2.96	0.52	44	3.37	0.52
Problem and solution text												
Top-level structure	59	4.37	0.85	54	5.06	0.79	48	4.24	0.81	44	4.38	0.89
Competence	59	3.65	0.70	54	5.07	0.91	48	3.64	0.64	44	4.43	0.94
Student level												
GSRT	1,222	36.07	11.53	1,131	40.02	11.66	887	35.46	11.51	893	37.42	13.47
Short comparison text												
Signaling test	1,415	13.83	8.32	1,415	15.80	9.46	1,074	12.81	8.32	1,074	13.17	9.43
Top-level structure	1,211	5.45	1.92	1,084	6.63	2.00	884	5.26	1.95	788	3.40	2.03
No. of issues	1,211	1.69	1.50	1,084	2.12	1.50	884	1.55	1.45	788	1.76	1.31
Competence	1,211	4.42	2.38	1,084	5.27	2.25	884	4.19	2.36	788	4.79	2.26
Long comparison text												
Top-level structure	1,214	2.72	1.59	1,082	4.70	2.22	884	2.68	1.57	787	4.13	2.04
# issues	1,214	0.22	0.57	1,082	0.70	0.98	884	0.18	0.48	787	0.47	0.78
Competence	1,214	1.51	1.14	1,082	2.81	1.67	884	1.46	1.02	787	2.42	1.53
Main idea												
Top-level structure	1,214	3.79	1.52	1,085	4.92	2.13	884	3.77	1.60	788	4.07	1.46
No. of issues	1,214	0.17	0.56	1,085	0.79	1.49	884	0.19	.54	788	0.22	0.61
Competence	1,214	3.01	1.52	1,085	3.85	1.43	884	2.96	1.56	788	3.40	1.37
Problem and solution text												
Top-level structure	1,214	4.43	2.50	1,081	5.14	2.42	884	4.25	2.48	787	4.44	2.32
Competence	1,214	3.69	2.08	1,081	5.13	2.59	884	3.67	2.03	787	4.50	2.54

Note. GSRT = Gray Silent Reading Test; ITSS = Intelligent Tutoring System for the Structure Strategy.

ifying the Level-1 coefficients for the reading pretests as a function of treatment to examine the question of whether the effect of ITSS on reading comprehension depends on students' initial reading level (M2). Similarly, a cross-level interaction between treatment and gender was added to M1 to test whether the effect of ITSS on reading comprehension differed between male and female students (M3). Moreover, a cross-level interaction between treatment and school locale was added to M1 to address the question of whether the effect of ITSS on reading comprehension varied across rural/suburban areas (M4). Statistically significant interactions were followed up by plotting the pattern of interaction. To test whether ITSS had different effects in different schools rather than having a common effect across all schools, we estimated variability of treatment effect across schools by modeling the Level-2 coefficients for treatment as random effects (M5). Statistically significant random treatment effects were followed up by estimating the 95% plausible value range of treatment effect among schools.

In addition, we estimated effect sizes of ITSS as compared with the control based on the main effect model (M1). Specifically, we computed the effect size as a standardized mean difference by dividing the adjusted (for pretest scores and other covariates) group mean difference by the (unadjusted) *pooled* within-treatment-group student-level standard deviation of the pretest scores. The use of pooled within-treatment-group student-level standard deviation to standardize effect estimate was recommended by What Works Clearinghouse (WWC, nodate, p.45).

Lastly, we examined simple Pearson correlations between the GSRT posttest and each of the indicators of system usage (average minutes used per week and total number of ITSS questions answered) for the ITSS group. A significant positive correlation would indicate that students who used the system more performed better on posttest. Moreover, a three-level regression model was conducted on GSRT posttest scores to examine relative effects of these two indicators of system usage after controlling for GSRT reading pretest scores.

Results

There was no statistically significant difference between ITSS and control groups on the pretests at the random assignment classroom level ($ps > .10$). This indicated that the ITSS and control classrooms were comparable in their reading level before the implementation of the experiment.

Class- and student-level simple descriptive statistics by treatment condition for GSRT and researcher-designed reading comprehension measures are presented in Table 1. Statistical test results of treatment effect from HLM analyses and effect sizes on GSRT, short comparison, long comparison, main idea, and problem and solution posttest scores are summarized in Table 2. HLM analyses (M0–M5) were conducted on each of the reading comprehension measures. However, for concern of space, we only present complete HLM results on the GSRT posttest (see Table 3). Effect estimates for ITSS presented in Table 3 were extracted from M1 for each of the outcome measures. Complete M1 estimates for all outcome measures are included in Table 3. Results are discussed by research questions.

Primary Research Question

To address the question of whether Grade 7 classrooms using the ITSS delivery of the structure strategy as a partial substitute for the standard language arts curriculum outperformed control classrooms on standardized and researcher-designed measures of reading comprehension, we used results from HLM Model 1 (see Table 3 M1 column). Students in ITSS classrooms on average scored 2.12 points (or 0.18 standard deviations) higher on GSRT adjusted posttest scores and 1.69 points (or 0.20 standard deviations) higher on short comparison Signaling posttest scores (see Table 2) than students in control classrooms holding reading pretest scores, gender, and school locale constant. These differences were statistically significant at $p < .05$. Adjusted posttest scores were also statistically significantly higher for students in ITSS classrooms than their control counterparts on all other researcher-designed reading comprehension measures (see Table 2), with effect sizes ranged from 0.15 on short comparison competence to 0.92 on main idea number of issues contrasted. The effect size of 0.18 on the standardized GSRT test was considered small, and the effect size of 0.92 on the main idea number of issues contrasted was considered large. Effect sizes on comparison top-level structure scores for recall from the short and long comparison texts as well as the main idea task were in the small medium range of 0.33 to 0.46.

Secondary Question 1

Results from model M2 provided an answer to the research question on whether the effect of ITSS on reading comprehension depended on students' initial reading level. For the number of issues contrasted in the main idea and long comparison text as well as main idea top-level structure on the posttest, the interaction between the student-level GSRT pretest and ITSS was significant at the .05 level. This indicated that the effect of ITSS, adjusted for other covariates in the model, varied depending on students' initial reading level as shown in Figures 6–8. The positive effect of ITSS on main idea number of issues, number of issues on the long comparison text, and main idea top-level structure on the posttest tended to be larger for students who had higher GSRT pretest scores.

There was also a statistically significant interaction between ITSS condition and student-level short comparison number of issues pretest on the short comparison number of issues posttest (see Figure 9). Figure 9 shows that the positive effect of ITSS on short comparison number of issues contrasted tended to increase as students' pretest scores increased. There were no statistically sig-

Table 2
Grade 7 Effect Sizes of ITSS on Reading Measures

Measure	Coefficient for ITSS (*SE*) from HLM[a]	Pooled student-level pretest standard deviation	Effect size
Gray Silent Reading Test	2.12*** (.48)	11.52	.18
Short comparison text			
Signaling test	1.69*** (.44)	8.32	.20
Top-level structure	.71*** (.12)	1.94	.37
No. of issues	.29* (.11)	1.48	.20
Competence	.36* (.12)	2.37	.15
Long comparison text			
Top-level structure	.52*** (.13)	1.58	.33
No. of issues	.17** (.06)	0.54	.31
Competence	.31** (.10)	1.09	.28
Main idea (short text)			
Top-level structure	.72*** (.10)	1.55	.46
No. of issues	.51*** (.08)	0.55	.92
Competence	.39*** (.06)	1.54	.25
Problem and solution text			
Top level structure	.59*** (.15)	2.49	.24
Competence	.52*** (.13)	2.05	.25

Note. Effect size = Adjusted difference between Intelligent Tutoring System for the Structure Strategy (ITSS; coded ½) and control (coded −½) groups divided by the student-level pooled standard deviation of pretest scores; HLM = hierarchical linear models.
[a] Estimates are extracted from Model 1; $df = 80$.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 3
HLM Results on Gray Silent Reading Test Posttest Scores for Grade 7

Variable	M0	M1	M2	M3	M4	M5
			Fixed effects			
Intercept	38.55*** (.62)	38.38*** (.38)	38.38*** (.38)	38.38*** (.38)	38.49*** (.33)	38.38*** (.38)
Rural		−.40 (.75)	−.40 (.75)	−.39 (.76)	−.58 (.72)	−.40 (.75)
Gray pretest		.42*** (.04)	.42*** (.04)	.42*** (.04)	.42*** (.04)	.42*** (.04)
Signaling pretest		.27*** (.05)	.27*** (.05)	.27*** (.05)	.27*** (.05)	.27*** (.05)
Class average Gray pretest		.65*** (.09)	.65*** (.09)	.64*** (.09)	.64*** (.09)	.64*** (.09)
Class average signaling pretest		.46***(.10)	.46*** (.10)	.46*** (.10)	.47*** (.11)	.46*** (.10)
Female		−.66 (.58)	−.65 (.58)	−.65 (.60)	−.67 (.58)	−.66 (.58)
ITSS		2.12*** (.48)	2.12*** (.48)	2.12*** (.48)	2.01*** (.46)	2.12*** (.48)
ITSS × Gray Pretest			−.04 (.08)			
ITSS × Signaling Pretest			−.02 (.10)			
ITSS × Female				−.21 (1.19)		
ITSS × Rural					1.43 (.95)	
			Random effects (variances of)			
Schools	2.48	0.98*	0.98*	0.99*	0.92*	1.00
Classrooms	20.04***	3.00**	3.01**	2.99**	2.98**	2.98
Students	136.13	108.04	108.00	108.05	108.03	108.05
ITSS						0.03
			Model fit statistics			
Deviance	15,720.69	14,099.04	14,098.23	14,098.99	14,098.05	14,099.06
Number of parameters	4	11	13	12	12	13

Note. M0 = unconditional model; M1 = main-effects model; M2 = interaction model with reading pretests; M3 = interaction model with gender; M4 = interaction model with school locale; M5 = random treatment-effects model; HLM = hierarchical linear models; Intelligent Tutoring System for the Structure Strategy.
* $p < .05$. ** $p < .01$. *** $p < .001$.

nificant interaction effects on other researcher-designed reading measures.

Secondary Questions 2 and 3

Models M3 and M4, respectively, addressed the research questions of whether the effect of ITSS on reading comprehension differed between male and female students and whether it varied across rural versus suburban areas. There was a statistically significant interaction between ITSS and gender on main idea number of issues contrasted and main idea top-level structure posttest scores. Figure 10 shows the similar pattern of interaction that the positive difference between ITSS and control groups on adjusted posttest scores for the number of issues compared (and main idea top-level structure) was slightly larger for female than for male students. The effect of ITSS did not appear to vary as a function of gender or school locale on any of the other reading outcomes that we examined. Holding reading pretest scores, research condition, and proportion of female students constant, suburban schools on average scored slightly higher than rural schools on main idea number of issues (0.22 point, $p < .05$), main idea top-level

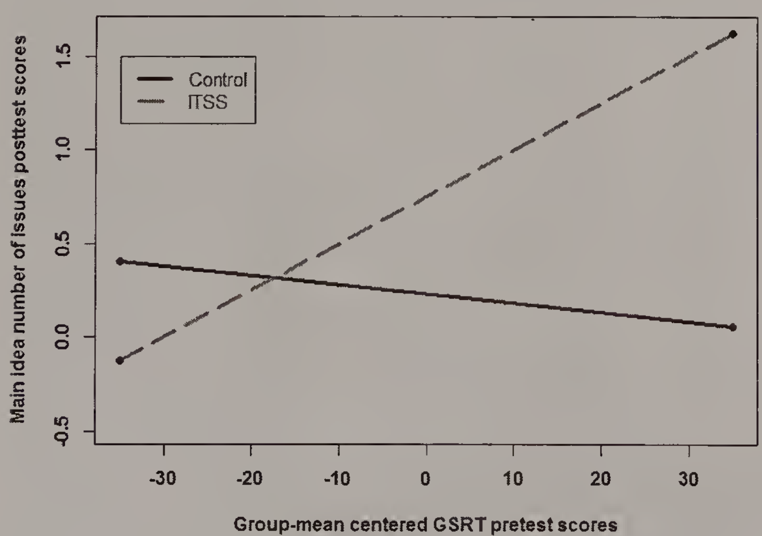


Figure 6. Interaction between experimental condition and Gray Silent Reading Test (GSRT) pretest level on main idea number of issues scores.

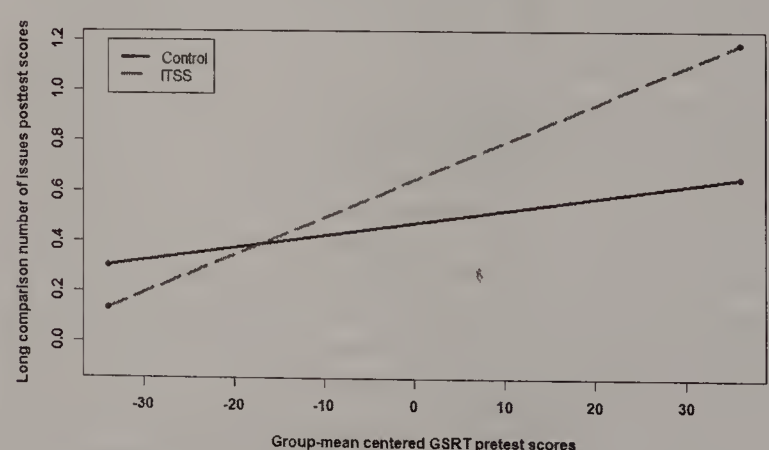


Figure 7. Interaction between experimental condition and Gray Silent Reading Test (GSRT) pretest level on long comparison text number of issues scores.

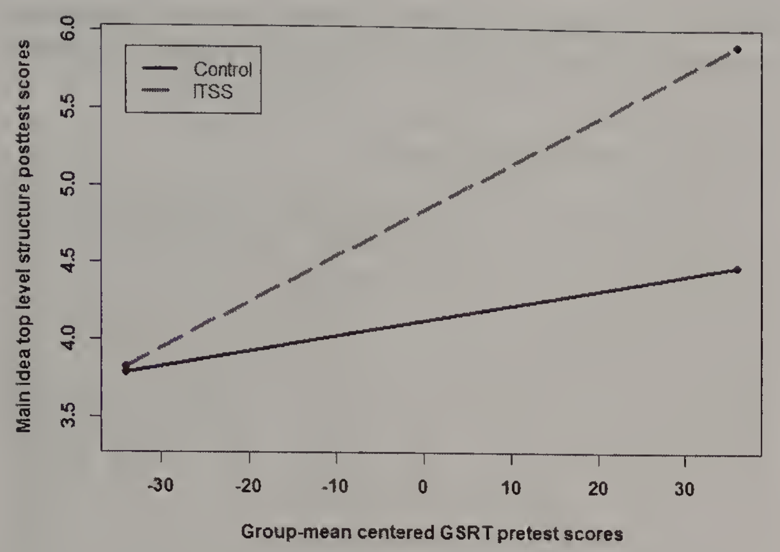


Figure 8. Interaction between experimental condition and Gray Silent Reading Test (GSRT) pretest level on main idea top-level structure scores.

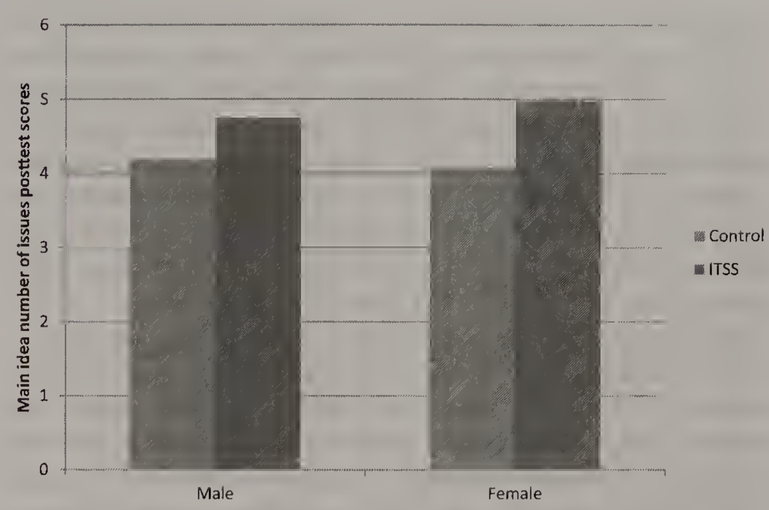


Figure 10. Interaction between experimental condition and gender on main idea number of issues scores.

structure (0.27 points, $p < .05$), and long comparison number of issues (0.13 points, $p < .05$). Holding reading pretest scores, research condition, and school locale constant, female students on average scored somewhat higher than male students on main idea competence (0.17 points, $p < .01$), long comparison number of issues (0.06 points, $p < .05$), long comparison competence (0.17 points, $p < .01$), long comparison top-level structure (0.23 points, $p < .01$), problem and solution competence (0.29 points, $p < .01$), problem and solution top-level structure (0.33 points, $p < .01$), short comparison number of issues (0.26 points, $p < .01$), short comparison competence (0.34 points, $p < .01$), short comparison top-level structure (0.23 points, $p < .001$), and signaling (0.91 points, $p < .01$). Students' gender and schools' locale did not seem to make a significant difference on the other posttest reading scores after pretest scores were controlled.

Secondary Question 4

The HLM model M5 addressed the question of whether the effect of ITSS on reading comprehension varied across schools. The estimated variance of adjusted ITSS effects across schools on the GSRT posttest and all researcher-designed reading measures

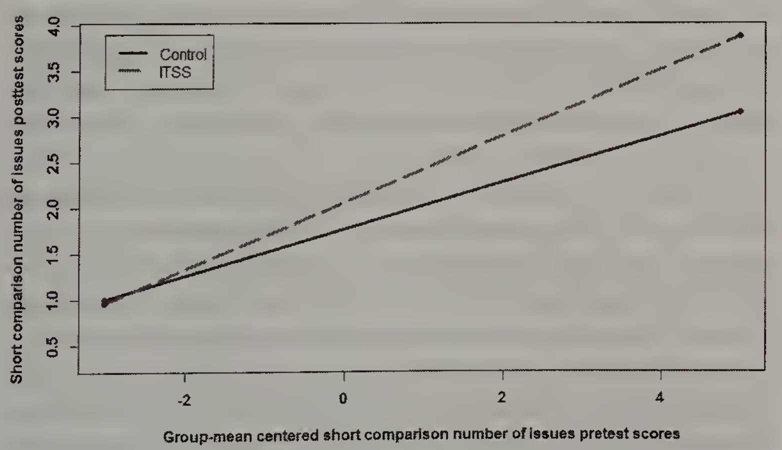


Figure 9. Interaction between experimental condition and short comparison number of issues pretest level on short comparison number of issues posttest scores.

was not statistically significantly different from zero at the .05 level. Difference in deviance between the random ITSS effect model (M5) and the fixed ITSS effect model (M1) was also not statistically significant on these measures (except main idea number of issues at $p < .05$ without correction for the number of tests). In other words, there was not sufficient evidence to suggest that adjusted ITSS effects (for the covariates) on the GSRT standardized test and researcher-designed measures differed significantly across schools. Therefore, the more parsimonious fixed-effects model was preferred.

Secondary Question 5

Finally, Pearson correlations between GSRT reading posttest scores and system usage measures were calculated to address the question of whether students who used the ITSS system for more time and who answered more questions performed better on the posttest. Average number of minutes used per week was not significantly related to GSRT posttest scores. However, the total number of questions answered demonstrated a positive and statistically significant correlation with GSRT ($r = .19$). Results from the three-level regression model also suggested that the total number of questions answered significantly predicted GSRT posttest scores above and over GSRT pretest and average number of minutes used per week ($b = .028$, $SE = .006$, $z = 4.72$, $p < .001$). In contrast, average number of minutes used per week became a negative predictor of GSRT posttest when both GSRT pretest and total number of questions answered were controlled for ($b = -.111$, $SE = .046$, $z = -2.41$, $p < .05$). These analysis results suggested that sheer time usage may not be a good indicator of fidelity. Students using extra time could be gaming the system rather than working on the lessons. The actual number of questions answered appeared to be a better indicator of fidelity as it indicated students' effort to learn the lessons. Questions refer to the requests I.T. made of the students within ITSS (e.g., write a main idea, write a recall, what is the cause?). As such, number of questions answered can be seen as a measure of student engagement within ITSS lessons.

As a sensitivity analysis, we also reanalyzed the fixed-effect model (M1) by adding affective pretest scores (computer attitudes,

reading self-concept, learning self-efficacy, and structure strategy self-efficacy) as covariates for the GSRT. Magnitudes of the adjusted ITSS effects on the standardized reading comprehension (GSRT) outcome measure remained about the same. Hence, detailed results of this analysis are not presented to conserve space. However, it might be worth noting that both student- and classroom- level learning self-efficacy and reading self-concept pretest scores were significant predictors ($p < .05$) for GSRT posttest scores holding reading pretest scores constant.

In summary, ITSS appeared to have a nontrivial positive influence on reading comprehension outcome measures above and over increases that could be predicted by students' initial reading and affective levels. The positive effect of ITSS also seemed to be stronger for students with higher reading pretest levels on several researcher-designed measures (i.e., main idea number of issues, long comparison number of issues, main idea top-level structure, and short comparison number of issues). Moreover, the effect of ITSS was somewhat larger for female students than for male students on the main idea number of issues and main idea top-level structure tests.

Summary of Classroom Observations and Computer Logs

Students in the intervention classrooms used the ITSS system for approximately 29 min each week for 22 weeks. They completed approximately 27 lessons on average. Most students completed instruction about the comparison, problem and solution, and cause and effect text structures. They also completed lessons on nested text structures.

The research team compiled weekly emails from the teacher-aides supporting the ITSS delivery and also summarized observations conducted in classrooms. Classroom observers were trained using video-taped segments created for the training. After training observers interrater reliability was 95%. Over 80% of the teachers did not participate in the computer lab time when ITSS was being used by the students with the support of the teacher-aides. Classroom observations focused on instructional foci and classroom organization. Observations showed that over 92% of teachers concentrated on literature and focused more on critiques, writing, and discussions. Less emphasis was paid to explicit text structure instruction and more focus was on implied text structure use (e.g., comparison of literary attributes). Due to budget and time limitations the team was unable to conduct observations in other content area classrooms, such as science.

Discussion

The purpose of this study was to examine the impact of ITSS used as a partial substitute for the language arts curriculum on 7th-grade students' reading comprehension. ITSS is a web-based intelligent tutoring system designed to teach 7th graders how to use the text structure strategy to read and comprehend expository texts by selecting and encoding strategic memory, summarizing, inferring, elaborating, and monitoring comprehension. The results showed that the text structure strategy delivered via the web-based ITSS had small but meaningful effects (.18) on the standardized reading comprehension (distal measure) and moderate to large effects on the proximal and distal researcher-designed measures of text structure competence, knowledge (i.e., signaling), and summaries (e.g., effect size of .91 on the main idea number of issues).

Research Findings in Context

The GSRT provides a sound distal measure of transfer of text structure knowledge to a standardized test. The GSRT measure contains inference and elaboration questions and the results from this study present a link from the text structure instruction to a standardized measure of reading comprehension. The effects on the proximal measures about constructing a strong main idea and free-recall tasks were larger. The signaling word task was slightly more distal than the main idea and recall tasks because students learned how to click on a signaling word in a well-signaled passage during ITSS instruction, but filled in blanks for the Signaling word dependent measure. Students also found it difficult to understand that multiple words can be placed in one blank (e.g., the same as). Thus, the effect on the signaling word task was not as strong as those for main idea and number of issues contrasted.

Overall in this study, effect sizes and outcomes were similar to those in recently published studies about ITSS at lower grade levels (Wijekumar et al., 2014, 2012) and other reading comprehension interventions not focusing primarily on text structure instruction (Cantrell et al., 2011; Slavin et al., 2008; Slavin, Chamberlain, Daniels, & Madden, 2009). The effect size on the GSRT was smaller than the results at fifth grade (Wijekumar et al., 2014), but stronger than for the fourth grade (Wijekumar et al., 2012). Cantrell et al. (2011) found improvements with 6th graders, but not 9th graders. Findings from the current study may be showing similar patterns of developmental challenges related to middle school students and the possibility that these students are developing poor habits that are difficult to change. At fifth grade, children showed malleability in cognitive processes and were receptive to interventions (Wijekumar et al., 2014). Findings from the current study and Cantrell et al. may signal a critical window of opportunity in upper elementary school for interventions to help students' improve their reading comprehension. These results may also be influenced by the schools choosing to provide their standard instruction to the poor readers and opting not to have most of them receive ITSS instruction. As noted in the introduction reading comprehension may be affected by the reader, text, and task variables. Students participating in this study may be affected by any one or more factors related to these areas. Further extensive qualitative and quantitative analyses of the ITSS computer logs may provide insight into the tasks that students completed and how students' online work within these tasks and subtasks affected the reading comprehension outcomes.

Results from this study were more robust than the studies on middle-grade reading comprehension with computer assisted instruction reviewed by Slavin et al. (2008), which showed a weighted mean effect size of +.10. The web-based ITSS appears to provide sound delivery, interactions, and learning environment for teaching the text structure strategy based on the weekly reports submitted by the teacher aides managing the ITSS rollout. The system has also shown stability in scaling up to larger numbers of users and is able to provide a meaningful, consistent, high quality alternative to relying solely on teacher delivery of instruction about the text structure strategy.

Students who scored at the highest levels at pretest showed the largest gains through using the ITSS. This finding was similar to results from a recent study on an intervention called the Reading Edge (Slavin et al., 2009) where similar interaction patterns for improvements were found for children reading below, at, or higher grade levels. However, the interaction pattern was different from findings using ITSS with students in Grades 4 and 5 (Wijekumar et al., 2014, 2012) and results from the smaller study with 7th-grade learners (Meyer et al., 2010). In these studies lower performing readers at the pretest (e.g., the GSRT) made greater gains after ITSS than stronger readers at pretest. The schedule limitations and pull-out instruction for special education in the middle schools may have contributed to this result. It was observed that students experiencing persistent reading difficulties in Grade 7 were receiving pull-out instruction during the ITSS times and thus missed receiving this intervention. Classroom observations showed that most teachers did not use the text structure strategy in the language arts classrooms. Teacher roles, their knowledge about text structure, and their fidelity of implementation with respect to consistency of instructions for the learners also may have contributed to the findings and should be carefully monitored in future studies.

The findings with 7th graders showed that most students could still benefit from text structure strategy instruction. They had not mastered the use of the text structure strategy by this grade level, and better readers clearly could benefit from the ITSS instruction as seen by organization of written recalls, generation of appropriate signaling words, quality of main ideas constructed, and increased ability to identify and contrast issues across paragraphs about different subtopics (i.e., different creatures of the same species in the shorter texts or different stone formations in the longer texts). It is unknown whether the greater jumps in performance after ITSS with below grade-level readers in the lower grades than in 7th grade resulted from more severe reading problems compounded by more years of failure or simply less opportunity to work in ITSS due to conflicts with pull out, remediation programs.

It is interesting to note that in the large randomized control trials with ITSS across 4th, 5th, and 7th grades, interactions between ITSS effects and gender varied from greater gains in males' ability to write good comparison mains ideas at Grade 4, to no interactions at Grade 5, and onto larger gains after ITSS for 7th-grade females than males on most experimenter-designed measures, but not the GSRT. This latter finding is compatible with Halpern's (2006) review that showed females to perform better than males when written responses are required rather than multiple-choice formats. Over all tested grades there were no interactions with ITSS and gender for posttest scores on the standardized, multiple-choice GSRT. Halpern (2006) also noted that males tend to comprise a greater proportion of students identified with severe to mild reading problems than females. In fourth grade, the lagging development-related reading skills for males in writing a main idea may have been particularly boosted by the heavily scaffolded ITSS instruction for writing a strong two-sentence main idea, a short writing task. In seventh grade, the greater gains for females may be due to a combination of dependent measures favoring writing tasks in which they can excel after text structure strategy instruction and less females with reading difficulties, which would result in fewer females missing ITSS due to pull out remediation sessions.

A number of factors that may have affected students' responsiveness to the intervention include students getting conflicting instruction from the teacher (vs. ITSS), little to no application of the text structure strategy in the language arts and/or content area classrooms, selections of texts, and the age and developmental level of the students. Classroom observations showed that teachers rarely spent time with students during the computer lab time when ITSS was implemented and relied on the teacher aides to monitor the class. Further, teacher surveys administered at the professional development session prior to the study showed that over 82% did not use text structure as part of the 7th-grade language arts curricula and none of them knew about the text structure strategy. Schools were reluctant to include content area classroom teachers (e.g., earth science) in the professional development due to the time commitment, and the research project did not have funds to conduct observations in those classrooms to document any text structure use in the content area classes.

A review of student responses presented some evidence about prior knowledge and practices impeding in the learning of the text structure strategy to improve reading comprehension. For example, one 7th-grade student wrote, "article compares two penguins but we should not notice differences, they are all the same." During the pretest and posttests students in two rural schools engaged in disruptive behaviors (e.g., excessive talking, running around the classroom). Teachers noted that the students had "given up" on education and were likely to drop out before entering high school.

Theoretical Implications

At the outset we compared the construction-integration and landscape models to the text structure model of reading comprehension, and we also compared reading comprehension interventions to the text structure strategy approach. Based on previous studies on the text structure strategy (e.g., Meyer et al., 1980), we reported that text structures and their direct and indirect scaffolds support the construction and integration of strategic memory from text. This strategic memory may be a representation of a coherent situation model identified in the construction-integration model of reading comprehension. Results from the current study provide further evidence in support of the text structure strategy in constructing strategic memories as evidenced by the 7th graders in ITSS producing stronger main ideas, organizing the main ideas using the centrality of connections, and utilizing the strategy when reading and responding to questions in a standardized test.

Practical Implications

A review of classroom observations and textbooks conducted by Wijekumar et al. (2013) showed that reading comprehension instruction at Grades 4 and 5 placed text structure as an independent and separate activity from summarizing, inferring, elaboration, and comprehension monitoring. At the 7th-grade level, observations showed there was even less emphasis on text structure and content area texts. Based on the accumulating evidence about the text structure strategy (current study; Wijekumar et al., 2014, 2012), students may benefit from reorganizing instruction to align with the text structure strategy and place text structure as the organizing framework for reading comprehension activities such as summa-

rizing, generating inferences, elaborating, and monitoring comprehension.

Results from the correlational analyses between the GSRT reading posttest scores and time on ITSS and the GSRT posttest scores and total questions answered also may provide insight into the use of web-based learning environments. These results suggested that students who focused their efforts on answering more questions about signaling words, writing main ideas and recall, and others ITSS tasks showed better performance on the GSRT posttest. Students spending extra time in this study may have been experiencing challenges in interacting with the system or gaming the system. Designers of computer-based interventions may take note and try to find approaches to encourage learners to actively engage in the practicc lessons and their performance tasks.

Limitations

The findings from this study may be generalizable to the extent that the populations of interest are similar to the sample studied here. This study used a volunteer sample of schools that was randomly assigned to the research conditions. A description of the sample is provided for researchers and practitioners to guide their interpretation of the results. Further research with different populations of students is necessary to extend these findings and examine broader generalizability. The participating sample did not include 7th-grade students who were receiving pull-out instruction and further research needs to be conducted with those special populations in the future.

Future Directions

Future research studies should also focus on the role of the teacher and stronger teacher professional development to support consistency of instruction so that students may learn the text structures presented in ITSS and receive consistent instructional support from the teacher. Further support to infuse text structure into the content area classrooms may improve the likelihood that students will see the utility of using the approach and reap the full benefits of strategic memory in the content area classrooms.

References

Alexander, P. A. (2005). The path to competence: A lifespan developmental perspective on reading. *Journal of Literacy Research, 37,* 413–436. http://dx.doi.org/10.1207/s15548430jlr3704_1

Allington, R. L. (2011). Reading intervention in the middle grades. *Voices from the Middle, 19,* 10–16. Retrieved from http://www.ncte.org.lib-ezproxy.tamu.edu:2048/journals

Biancarosa, G., & Snow, C. E. (2006). *Reading next—A vision for action and research in middle and high school literacy: A report to Carnegie Corporation of New York* (2nd ed.). Washington, DC: Alliance for Excellent Education.

Bohn-Gettler, C. M., & Kendeou, P. (2014). The interplay of reader goals, working memory, and text structure during reading. *Contemporary Educational Psychology, 39,* 206–219. http://dx.doi.org/10.1016/j.cedpsych.2014.05.003

Caccamise, D., Friend, A., Groneman, C., Littrell-Baez, M. K., & Kintsch, E. (2014). Teaching struggling middle school readers to comprehend informational text. In J. L. Polman, E. A. Kyza, D. K. O'Neill, I. Tabak, W. R. Penuel, A. S. Jurow, . . . L. D'Amico (Eds.), *Learning and becoming in practice: The International Conference of the Learning Sciences* (Vol. 2, pp. 1002–1006). Boulder, CO: International Society of the Learning Sciences; http://www.colorado.edu/ics/sites/default/files/attached-files/caccamise_et_al._2014_icls_proceedings1.pdf

Cantrell, S. C., Almasi, J. F., Carter, J. C., Rintauma, M., & Madden, A. (2010) The impact of a strategy-based intervention on the comprehension and strategy use of struggling adolescent readers. *Journal of Educational Psychology, 102,* 257–280

Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S., . . . Schatschneider, C. (2011). Testing the impact of Child Characteristics × Instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly, 46,* 189–221.

Duke, N. K. (2010). The real-world reading and writing U.S. children need. *Phi Delta Kappan, 91,* 68–71. http://dx.doi.org/10.1177/003172171009100517

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of Educational Research, 79,* 262–300. http://dx.doi.org/10.3102/0034654308325998

Foresman, S. (2007). *Reading street.* Glenview, IL: Pearson/Scott Foresman.

Freudenrich, C. (2000). *How landfills work.* Retrieved from http://science.howstuffworks.com/environmental/green-science/landfill.htm

Gernsbacher, M. A. (1996). The structure-building framework: What it is, what it might also be, and why. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 289–312). Mahwah, NJ: Erlbaum.

Gewertz, C. (2011). Common-Core writers craft curriculum criteria. *Education Week, 30,* 1–5.

Goldman, S. R., Varma, S., & Cote, N. (1996). Extending capacity-constrained construction integration: Toward "smarter" and flexible models of text comprehension. In B. K. Britton & A. C. Grasesser (Eds.), *Models of understanding text* (pp. 73–113). Hillsdale, NJ: Erlbaum.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530

Guthrie, J. T., & Davis, M. H. (2003). Motivating struggling readers in middle school through an engagement model of classroom practice. *Reading & Writing Quarterly, 19,* 59–85. http://dx.doi.org/10.1080/10573560308203

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* London, England: Longman.

Halpern, D. F. (2006). Assessing gender gaps in learning and academic achievement. In E. Anderman, P. H. Winne, P. A. Alexander, & L. Corno (Eds.), *Handbook of educational psychology* (2nd ed., pp. 635–654). Mahwah, NJ: Erlbaum.

Hammann, L. A. (2000). *An investigation of instruction in summarizing and text structure for compare-contrast writing* (Unpublished doctoral dissertation). The Pennsylvania State University, University Park, PA

Hebert, M., Bohaty, J. J., Nelson, J. R., & Brown, J. A. (2016). The effects of text structure instruction on informational text comprehension: A meta-analysis. *Journal of Educational Psychology, 108,* 609–629.

Kim, A. H., Vaughn, S., Klingner, J. K., Woodruff, A. L., Reutebuch, C. K., & Kouzekanani, K. (2006). Improving the reading comprehension of middle school students with disabilities through computer-assisted collaborative strategic reading. *Remedial and Special Education, 27,* 235–249. http://dx.doi.org/10.1177/07419325060270040401

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* New York, NY: Cambridge University Press.

Kintsch, W., & Kintsch, E. (2004). Comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71–92). Mahwah, NJ: Erlbaum.

Mallette, M. H., Duke, N. K., Strachan, S. L., Waldron, C. H., & Watanabe,

L. M. (2013). Synergy in literacy research methodology. In D. E. Alvermann, N. J. Unrau, & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (6th ed., pp. 91–128). Newark, DE: International Reading Association. http://dx.doi.org/10.1598/0710.03

Mason, L. H. (2004). Explicit self-regulated strategy development versus reciprocal questioning: Effects on expository reading comprehension among struggling readers. *Journal of Educational Psychology, 96,* 283–296.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38,* 1–30. http://dx.doi.org/10.1207/s15326950dp3801_1

McNamara, D. S., O'Reilly, T. P., Best, R. M., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research, 34,* 147–171. http://dx.doi.org/10.2190/1RU5-HDTJ-A5C8-JVWE

Meyer, B. J. F. (1975). *The organization of prose and its effects on memory.* Amsterdam, the Netherlands: North-Holland.

Meyer, B. J. F. (1984). Organizational aspects of text: Effects on reading comprehension and applications for the classroom. In J. Flood (Ed.), *Promoting reading comprehension* (pp. 113–138). Newark, DE: International Reading Association.

Meyer, B. J. F. (2003). Text coherence and readability. *Topics in Language Disorders, 23,* 204–224. http://dx.doi.org/10.1097/00011363-200307000-00007

Meyer, B. J. F., Brandt, D. M., & Bluth, G. J. (1980). Use of the top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading Research Quarterly, 16,* 72–103. http://dx.doi.org/10.2307/747349

Meyer, B. J. F., & Freedle, R. O. (1984). Effects of discourse type on recall. *American Educational Research Journal, 21,* 121–143. http://dx.doi.org/10.3102/00028312021001121

Meyer, B. J. F., & Poon, L. W. (2001). Effects of the structure strategy training and signaling on recall of text. *Journal of Educational Psychology, 93,* 141–159. http://dx.doi.org/10.1037/0022-0663.93.1.141

Meyer, B. J. F., & Wijekumar, K. (2007). A Web-based tutoring system for the structure strategy: Theoretical background, design, and findings. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 347–375). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Meyer, B. J. F., Ray, M. N., & Middlemiss, W. (2012). Children's use of comparative text signals: The relationship between age and comprehension ability. *Discours Revue de linguistique, psycholinguistique et informatique, 10,* 1–25. http://dx.doi.org/10.4000/discours.8637

Meyer, B. J. F., & Rice, G. E. (1989). Prose processing in adulthood: The text, the reader, and the task. In L. W. Poon, D. C. Rubin, & B. A. Wilson (Eds.), *Everyday cognition in adulthood and later life* (pp. 157–194). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511759390.013

Meyer, B. J. F., Wijekumar, K., Middlemiss, W., Higley, K., Lei, P., Meier, C., & Spielvogel, J. (2010). Web-based tutoring of the structure strategy with or without elaborated feedback or choice for fifth- and seventh-grade readers. *Reading Research Quarterly, 45,* 62–92. http://dx.doi.org/10.1598/RRQ.45.1.4

National Assessment of Educational Progress. (2013). *The nation's report card.* Retrieved from http://nationsreportcard.gov/reading_math_2013/#/

Pearson, P. D., & Hiebert, E. H. (2015). *Research-based practices for teaching Common Core literacy.* New York, NY: Teachers College Press.

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 545–561). Mahwah, NJ: Erlbaum.

ProEd. (2015). *GRST: Gray Silent Reading Tests.* Available at http://www.proedinc.com/customer/productView.aspx?ID=1743

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1993). Higher order instructional goals in secondary schools: Class, teacher, and school influences. *American Educational Research Journal, 30,* 523–553. http://dx.doi.org/10.3102/00028312030003523

Raudenbush, S. W., & Bryk, A. S. (2002). *Hieararchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: SAGE publications.

Reardon, S. E., Valentino, R. A., & Shores, K. A. (2012). Patterns of literacy among U.S. students. *The Future of Children, 22,* 17–37. http://dx.doi.org/10.1353/foc.2012.0015

Schellings, G. L. M., & Broekkamp, H. (2011). Signaling task awareness in think-aloud protocols from students selecting relevant information from text. *Metacognition and Learning, 6,* 65–82. http://dx.doi.org/10.1007/s11409-010-9067-z

Slavin, R. E., Chamberlain, A., Daniels, C., & Madden, N. A. (2009). The Reading Edge: A randomized evaluation of a middle school cooperative reading program. *Effective Education, 1,* 13–26. http://dx.doi.org/10.1080/19415530903043631

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effetive reading programs for middle and high schools. A best-evidence synthesis. *Reading Research Quarterly, 43,* 290–322. http://dx.doi.org/10.1598/RRQ.43.3.4

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360–407. http://dx.doi.org/10.1598/RRQ.21.4.1

Stine-Morrow, E. A. L., Gagne, D. D., Morrow, D. G., & DeWall, B. H. (2004). Age differences in rereading. *Memory & Cognition, 32,* 696–710. http://dx.doi.org/10.3758/BF03195860

Taboada, A., Tonks, S. M., Wigfield, A., & Guthrie, J. (2009). Effects of motivational and cognitive variables on reading comprehension. *Reading and Writing: An Interdisciplinary Journal, 22,* 85–106.

Taylor, B. M., Graves, M. F., & van den Broek, P. W. (Eds.). (2000). *Reading for meaning: Fostering comprehension in the middle grades.* New York, NY: Teachers College Press.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York, NY: Academic Press.

Vaugh, S., Wanzek, J., Swanson, E., & Roberts, G. (2015, July). *Efficacy of comprehension and content acquisition for middle school students who are and are not English language learners.* Paper presented at the Annual Conference of the Society for the Scientific Study of Reading, Waikaloa, HI.

Vaughn, S. (2015, July). *Improving content knowledge and comprehension for English language learners: Findings from two randomized control trials.* Paper presented at the Annual Conference of the Society for the Scientific Study of Reading, Waikaloa, HI.

Voss, J. F., & Silfies, L. N. (1996). Learning from history text: The interaction of knowledge and comprehension skill with text structure. *Cognition and Instruction, 14,* 45–68. http://dx.doi.org/10.1207/s1532690xci1401_2

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22,* 333–362.

Wheldall, K. (2005). When will we ever learn? *Educational Psychology, 25,* 573–584. http://dx.doi.org/10.1080/01443410500344639

Wiederholt, J. L., & Blalock, G. (2000). *Gray Silent Reading Tests (GSRT).* Austin, TX: Pro-Ed.

Wijekumar, K. (2014a). *The comparison text structure for adults and kids* [Video]. http://youtu.be/d_ZL0yEeUac

Wijekumar, K. (2014b). *The problem and solution and cause and effect text structures for adults and kids* [Video]. http://youtu.be/gEEXvMPMU2k

Wijekumar, K., Meyer, B. J. F., Harris, K., Graham, S., & Beerwinkle, A. (2016). Comparing learning outcomes and implementation factors from student-managed vs. teacher-managed intelligent tutoring systems. In R. K. Atkinson (Ed.), *Intelligent tutoring systems: Structure, applications, and challenges* (pp. 175–200). Hauppauge, NY: Nova Science.

Wijekumar, K., Meyer, B. J. F., & Lei, P. (2012). Large-scale randomized controlled trial with 4th graders using intelligent tutoring of the structure strategy to improve nonfiction reading comprehension. *Educational Technology Research and Development, 60,* 987–1013. http://dx.doi.org/10.1007/s11423-012-9263-4

Wijekumar, K. K., Meyer, B. J. F., & Lei, P. (2013). High-fidelity implementation of web-based intelligent tutoring system improves fourth and fifth graders content area reading comprehension. *Computers & Education, 68,* 366–379. http://dx.doi.org/10.1016/j.compedu.2013.05.021

Wijekumar, K., Meyer, B. J. F., Lei, P., Lin, Y., Johnson, L. A., Spielvogel, J. A., . . . Cook, M. (2014). Multisite randomized controlled trial examining intelligent tutoring of structure strategy for fifth-grade. *Journal of Research on Educational Effectiveness, 7,* 331–357. http://dx.doi.org/10.1080/19345747.2013.853333

Yeari, M., & van den Broek, P. (2011). A cognitive account of discourse understanding and discourse interpretation: The landscape model of reading. *Discourse Studies, 13,* 635–643. http://dx.doi.org/10.1177/1461445611412748

Received December 28, 2015
Revision received October 31, 2016
Accepted November 1, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 6, 761–781

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000180

Examining the Impact of Inference Instruction on the Literal and Inferential Comprehension of Skilled and Less Skilled Readers: A Meta-Analytic Review

Amy M. Elleman
Middle Tennessee State University

Inference ability is considered central to discourse processing and has been shown to be important across models of reading comprehension. To evaluate the impact of inference instruction, a meta-analysis of 25 inference studies in Grades K–12 was conducted. Results showed that inference instruction was effective for increasing students' general comprehension, $d = 0.58$, inferential comprehension, $d = 0.68$, and literal comprehension, $d = 0.28$. Although skilled and less skilled readers responded similarly on general and inference outcomes, less skilled readers benefited more on literal outcomes, $d = 0.97$, than skilled readers, $d = 0.06$. Findings suggest that students can increase their inference ability and that less skilled readers gain the extra benefit of increases in literal comprehension. Findings also suggest that instruction provided in small groups is beneficial for increasing readers' inferential understanding of text.

Keywords: inference, comprehension, instruction, meta-analysis, reading difficulties

Understanding the implied message is fundamental to any act of communication. Researchers in linguistics, cognitive psychology, and education consider inferential processing a central component in language comprehension and essential to reading comprehension (e.g., Bransford & Franks, 1971; Cain & Oakhill, 2007; Garnham & Oakhill, 2014; Kintsch & Kintsch, 2005; Singer, 1988). Rarely are texts fully explicit. If every idea had to be explicitly articulated in a text, the text would be lengthy and boring. Writers construct prose with the expectation that readers will fill in gaps using clues provided in the text or the reader's prior knowledge. To fully gain the intended meaning from a text, a reader must go beyond a surface level understanding to create a coherent mental representation, or situation model (Kintsch, 1988; van Dijk & Kintsch, 1983). At the heart of this process is inference generation, the process by which a reader integrates information within or across texts using his or her background knowledge to fill in information not explicitly stated (Kendeou, 2015; McNamara & Magliano, 2009).

This article was published Online First February 13, 2017.

This research was supported in part by Grant R305G050101 from the U.S. Department of Education, Institute of Education Sciences to Vanderbilt University. Statements do not reflect the position or policy of these agencies and no official endorsement by them should be inferred. I am grateful to Donald Compton, Doug Fuchs, Lynn Fuchs, and Joseph Jenkins for their review of the article and helpful suggestions. I am also grateful to Endia Lindo, Susan Amundrud, Summer Talbert, and Stacy Fields for their review of the article and help in coding the studies for interobserver agreement.

Correspondence concerning this article should be addressed to Amy M. Elleman, Literacy Studies Ph.D. Program, College of Education, Middle Tennessee State University, Murfreesboro, TN 37132. E-mail: amy.elleman@mtsu.edu

Theoretical Underpinnings

Inference generation has played a prominent role across leading theories and models of reading comprehension (for a review see Kendeou, McMaster, & Christ, 2016; McNamara & Magliano, 2009). Some models assume inference generation is based mostly on automatic memory retrieval processes (e.g., Albrecht & O'Brien, 1993), whereas others posit that some inferences are actively controlled by the reader (e.g., Graesser, Singer, & Trabasso, 1994; Trabasso & van den Broek, 1985; van den Broek et al., 2005). Many models consider the role of knowledge-based inferences in comprehension (Kintsch, 1988; Kendeou, Walsh, Smith, & O'Brien, 2014), whereas a few focus only on the impact of text-based inferences for maintaining coherence (e.g., Zwaan, Magliano, & Graesser, 1995). Despite these different perspectives, nearly all discourse comprehension models assume that readers establish coherence by generating text-based inferences to make sense of causal, temporal, and spatial relationships in discourse (McNamara & Magliano, 2009).

Kintsch's (1988) Construction-Integration (CI) model is considered the most comprehensive explanation of comprehension to date (McNamara & Magliano, 2009). The CI model assumes that memory representations are stored as nodes and links in an underlying connectionist architecture. In the construction phase of CI, information from the text and related knowledge from memory are automatically activated. In the subsequent integration process, activation spreads throughout the nodes and links settling on those concepts with more links to other concepts and greater activation while disregarding concepts with fewer links. These integrative processes occur iteratively across text (i.e., propositions, sentences, and passages) allowing the reader to continually update his or her mental representation during reading.

Another important aspect of the CI model is its consideration of related, but distinct levels of representation (McNamara & Magliano, 2009). The CI model posits that a reader's understand-

ing of the text is dependent on three levels of representation: the surface level, the textbase, and the situation model (Kintsch, 1988). The surface level represents memory for the specific words and phrases in the text. The surface level has not been explored in any depth by discourse researchers because it is thought to minimally impact comprehension (McNamara & Magliano, 2009). The textbase level refers to the explicit ideas expressed in the text and requires minimal inferential processing. The situation model, on the other hand, relies heavily on inference generation, as it requires the reader to integrate his or her prior knowledge with the information in the text. This integration leads to durable learning and likely facilitates the use of the acquired knowledge in new situations (McNamara, Kintsch, Songer, & Kintsch, 1996).

The Role of Inference in Reading Comprehension

While inferential processes are assumed to be a central aspect of comprehension in discourse process models, component models make no such assumption. Component models have been used to empirically test the relationship between inference and comprehension (Ahmed et al., 2016; Cromley & Azevedo, 2007; Cromley, Snyder-Hogan, & Luciw-Dubas, 2010). One such model, the Direct and Inferential Mediational Model (DIME) model has been used to directly test relationships among components thought to be important for reading comprehension (e.g., inference, knowledge, vocabulary, word reading, and strategies). In their first study using the DIME model, Cromley and Azevedo (2007) found that vocabulary and background knowledge made the largest contribution to reading comprehension followed by inference ability, word reading, and strategies for students in Grade 9. In a subsequent study with college students, the best-fitting model included direct effects for knowledge, vocabulary, and inference and indirect effects for vocabulary and knowledge through strategies and inference (Cromley et al., 2010). In a validation study using a larger and more diverse sample of students in Grades 7–12, Ahmed et al. (2016) found that vocabulary and knowledge had less of an impact on comprehension and that inference played a more prominent role than in previous studies.

Inference ability has been shown to be a unique predictor of reading comprehension at multiple developmental stages (e.g., Barth, Barnes, Francis, Vaughn, & York, 2015; Cain, Oakhill, Barnes, & Bryant, 2001; Cain, Oakhill, & Bryant, 2004). In a longitudinal study, Cain and Oakhill (2007) showed that inference skill for elementary age children was significantly related to comprehension, over and above IQ, word reading, and vocabulary knowledge two years later. In another study, Cain and Oakhill (1999) matched 7 to 8 year-old skilled and less skilled comprehenders with a group of 6 year-old children on general comprehension. Both the older skilled readers and younger matched readers were better at making text-connecting inferences than the older less skilled readers, despite that both the younger and older less skilled groups had the same general comprehension abilities. Based on these findings, Cain and Oakhill suggest that inference ability is not a by-product of comprehension, but should be considered a plausible cause of reading comprehension ability.

Individual Differences in Inference Ability

Oakhill and her colleagues have conducted numerous studies with poor comprehenders (i.e., students who are fluent and accurate at decoding, but poor at gaining meaning from text) to better understand the role of inference generation in comprehension. In a series of experiments, they found that good and poor comprehenders, matched on vocabulary and decoding, differ in their ability to make inferences at each level of textual discourse (i.e., word, sentence, and passage level; e.g., Cain & Oakhill, 1999; Oakhill, 1984; Yuill & Oakhill, 1991). One striking example is the inability of poor comprehenders to use simple linguistic devices that signal the cohesion of a text such as anaphoric references. Yuill and Oakhill presented sentences containing two types of referents to 7- and 8- year-olds. For example, children were presented with the sentence, "*Peter lent his coat to Sue because she was cold*" and were then asked, "Who was very cold, Peter or Sue?" (Yuill & Oakhill, 1991, p. 81). The less skilled comprehenders made errors even when gender cues were available, suggesting integrative problems at a basic level. Other studies have also demonstrated that skilled readers differ from less skilled readers in regards to inference generation (e.g., Barth et al., 2015; Bowyer-Crane & Snowling, 2005; Cain & Oakhill, 1999). Skilled readers have been shown to be more likely than less skilled readers to generate topic-related inferences during comprehension, to integrate words into text and with preceding context, to select homographs and answer inference questions in a logical manner (Kintsch & Mross, 1985; Long, Oppy, & Seely, 1994; Perfetti & Stafura, 2014; Wilson, 1979). Studies have also shown that readers' responsiveness to inference instruction differs based on reading ability, strategic ability, and level of prior knowledge (McGee & Johnson, 2003; McNamara, O'Reilly, Best, & Ozuru, 2006).

There are a number of hypotheses that attempt to explain why some children have difficulties making inferences from text. One hypothesis is that comprehension problems are because of differences in working memory capacity (Daneman & Carpenter, 1983). Working memory capacity refers to the amount of information that can be temporarily stored and manipulated during a task. Working memory capacity may differentially impact inferences required to maintain local coherence and those necessary for global coherence (Albrecht & O'Brien, 1993; Graesser et al., 1994). A reader must establish local coherence by connecting new textual information with the immediately preceding text. Global coherence, on the other hand, is established by making connections with information that is no longer available in short-term memory (STM). In this way, working memory functions as a buffer for the incoming ideas in a text. A larger working memory capacity promotes the probability that the necessary ideas in a text will be available for integration (Singer, Andruslak, Reisdorf, & Black, 1992). It is, therefore, not surprising that working memory capacity has been shown to explain individual differences in reading comprehension (Cain, Oakhill, & Lemmon, 2004; De Beni, Palladino, Pazzaglia, & Cornoldi, 1998; Seigneuric & Ehrlich, 2005; Swanson & Berninger, 1995).

Other researchers contend that the driving force behind inference generation is because of activation and use of prior knowledge (Kintsch, 1998; Kintsch & Kintsch, 2005; McNamara, O'Reilly, & de Vega, 2007). Kintsch and Kintsch (2005) asserted that most inferences are a matter of knowledge retrieval and theorized that greater activation because of well-connected memory structures facilitates retrieval and allow for the creation of a more coherent situation model, which in turn facilitates inference generation. It has been consistently shown that readers with higher

levels of knowledge have an advantage on comprehension and memory tasks compared to those with lower levels of knowledge (e.g., Chiesi, Spilich, & Voss, 1979; Kendeou & van den Broek, 2007; Recht & Leslie, 1988). It is also well established that readers who possess more knowledge of a domain area are better able to understand a text even when multiple inferences are required (Kendeou et al., 2014; McNamara et al., 1996; O'Reilly & McNamara, 2007). Alternatively, some researchers contend that it is not necessarily activation of prior knowledge that is the primary culprit for poor comprehension, but difficulty suppressing irrelevant knowledge (e.g., Gernsbacher & Faust, 1991; Rosen & Engle, 1997). Accounts of poor readers' illogical and haphazard answers to inferential comprehension questions support this account (e.g., Williams, 1993; Wilson, 1979). Knowledge proponents, however, counter that the inability to suppress information is not because of faulty suppression, but is instead because of poor comprehenders' lack of high quality knowledge networks that facilitate efficient retrieval of information (Kendeou et al., 2014; McNamara & McDaniel, 2004).

Knowledge is necessary for inference generation, but Barnes and colleagues demonstrated that other memory-based and integrative factors are also important in inference generation (Barnes, Dennis, & Haefele-Kalvaitis, 1996). To control for differences in background knowledge, Barnes et al. taught children a new knowledge base about a fictional world, Gan. Students then read a story and had to integrate facts about Gan with what was happening in a story. Although the children had the requisite knowledge necessary to make the appropriate inferences, they still had difficulty answering inferential questions about the story. In a subsequent study using the same story and similar methodology with good and poor comprehenders, Cain et al. (2001) found that even after controlling for students' background knowledge, poor comprehenders made fewer accurate inferences than good comprehenders. These findings suggest that issues with inference generation cannot be wholly accounted for by individual differences in background knowledge.

Although researchers disagree on the driving mechanism behind inference generation, most acknowledge the prominent role executive function (e.g., self-regulation, metacognition, attention) plays in higher order comprehension processes (Oakhill, Hartt, & Samols, 2005; Sesma, Mahone, Levine, Eason, & Cutting, 2009). Research has shown that young children and poor readers often fail to make inferences unless prompted to do so (e.g., Cain & Oakhill, 1999; Hannon & Daneman, 1998; Paris & Lindauer, 1976). Comprehension monitoring and strategy use are essential for inference generation. An effective reader must continually check his or her mental model against the text to determine if an inference is required. Once an inconsistency is detected, he or she must know which strategy will be most efficient at reconciling his or her mental model with the information in the text. Unfortunately, poor readers often fail to recognize their own comprehension breakdown and are less likely to use effective strategies compared to good readers (e.g., Long & Chong, 2001; Markman, 1979; Oakhill et al., 2005).

Inference Instruction

While researchers have made advances in our understanding of inferential processes, few studies have examined the opportunities students are given to learn inference skills in the classroom. The first observational study considering inference instruction practices in the classroom suggested that little time was spent encouraging inference generation (Guszak, 1967). Guzak found that only 20% of teachers' questions were inferential. Since then, only two studies have considered inference generation opportunities in classroom practice (Franks, Mulhern, & Schillinger, 1997; O'Flahavan, Hartman, & Pearson, 1989). O'Flahavan et al. conducted an observational study and found that more than half of the questions teachers used were inferential. Later, in a review of reading basals, Franks et al. found that although there were opportunities provided in the basals for students to make logical inferences, there were few instructions for teachers in how to *teach* such inferences.

Despite a lack of knowledge about opportunities and practices for learning inference skills in today's classroom context, there is research supporting the instruction of inference generation. Most of these inference interventions reflect the theoretical accounts considered in the literature. When children are taught to monitor their comprehension and use strategies to better understand text, their inference skills have been shown to improve (e.g., Dewitz, Carr, & Patberg, 1987; McNamara, 2004; McNamara et al., 2006; Yuill & Oakhill, 1988). Many researchers have found that teaching students to use self-generated elaborations enhances comprehension, including higher-order reasoning (King, 1994; King & Rosenshine, 1993; King, Staffieri, & Adelgais, 1998; McNamara et al., 2006; Spires & Donley, 1998; Stein et al., 1982). In addition to teaching students to use inference strategies and self-regulation, programs focused on teaching children to activate background knowledge have also been shown to be successful (e.g., Gordon, 1980; Hansen & Pearson, 1983).

One question that has not been addressed fully in the inference literature concerns the impact of inferential processing on the memory for literal content of a text. Memory research examining gist processing suggests that once a reader synthesizes the pertinent information from a context, his or her memory for the verbatim content begins to degrade (Brainerd & Reyna, 2005; Bransford & Franks, 1971). According to this view, if students are taught inference strategies to strengthen building a coherent situation model, their verbatim memory for the text content will weaken, possibly leading to poorer performance on literal questions. Alternatively, Kintsch's (1988) CI model suggests that active construction of the situation model can feed back into the proposition level and strengthen memory for the literal content. According to this view, students' memory for literal content should be enhanced after developing a coherent situation model. It is interesting that beyond Kintsch's CI model, other comprehension discourse models focus mainly on the development of the situation model and spend little time explicitly addressing the impact of the inference processing on memory of textbase representations (McNamara & Magliano, 2009). This lack of focus seems at odds with the comprehension intervention research in which researchers often test the impact of instruction on students' literal and inferential understanding of texts.

A separate, but related issue concerns the impact of exclusively teaching inference strategies to less skilled readers. Poor comprehenders not only demonstrate difficulties with inference (e.g., Barth et al., 2015; Cain et al., 2001) and gist processing (Weekes, Hamilton, Oakhill, & Holliday, 2008), but they have also been

shown to require more practice to learn the literal content of a new knowledge base (Cain et al., 2001). These deficits suggest that struggling readers may need additional strategies and support to help them understand at the textbase level of representation, before they can learn to effectively use inference strategies to build a coherent situation model.

Study Purpose

Decades of research have provided extensive knowledge regarding inference generation. The purpose of this meta-analytic review is to examine the impact of these instructional approaches on the reading comprehension of skilled and less skilled readers in Grades K–12. This review considers questions relevant to both research and practice by examining the relationship between inference and comprehension, the impact of inference instruction on different levels of text representation, the differential response of skilled and less skilled readers to inference instruction, and key factors associated with instructional effectiveness.

Theoretical models of discourse processing consider inference to be a central component of reading comprehension and findings from many empirical studies demonstrate a strong relationship between inference generation and reading comprehension. Intervention studies are uniquely designed to answer questions involving causality. We, therefore, reviewed all of the quasi and experimental studies that directly tested the impact of teaching inference on students' reading comprehension. We were also interested in how inference instruction impacts different levels of text representation, so we considered the impact of inference instruction on both literal and inferential outcomes. Skilled and less skilled readers differ on multiple dimensions in regards to inference generation and reading comprehension, so we also wanted to explore the ways they might differentially respond to inference instruction. In addition to reading ability, we were also interested in other participant characteristics (e.g., socioeconomic status [SES], grade level) that might be associated with instructional effectiveness.

Another primary goal of this review was to identify key factors associated with instructional effectiveness. Identification of these factors could help inform current instructional practices and assist in the development of more effective interventions in reading comprehension. We were especially interested in knowing if interventions focusing on components thought to be important to inference generation (e.g., activating background knowledge, integrating text information, and metacognitive strategy use) would be more effective than other types of instruction. We also considered other intervention characteristics (e.g., study length, type of text, type of inferences taught, group size, level of discussion, and incorporation of writing) that might be associated with outcomes.

Method

Study Inclusion Criteria

General study and intervention characteristics. To include as much evidence as possible, a wide range of publications published between 1950 and 2014 were eligible for review including journal articles, dissertations, and reports. These studies were included to cover as much of the extant literature while still being useful for informing modern pedagogical practices.

Only instructional methods focused on increasing students' comprehension by improving their ability to make inferences were eligible for inclusion. One of the purposes of this review was to consider the impact of instructional techniques intended to improve inferential skills, not general comprehension instruction that includes strategies intended to improve literal and inferential understanding. Another purpose of the review was to consider what happens to readers' literal understanding of the text after receiving instruction focused on the inferential content. To evaluate this as clearly as possible, studies that focused primarily on teaching inferences were included, and studies teaching literal strategies along with inferential strategies were excluded. To be included in this review, components such as summarizing had to be a minimal part of the intervention (comprising less than 20% of the intervention). In addition, short experimental studies conducted to understand the cognitive process of making inferences, but not focused on instructional techniques that could be used in the classroom, were not included.

Outcome measures. Each measure was identified as inferential (i.e., questions asking for information not explicitly stated in the text), literal (i.e., questions asking for information explicitly stated in text), or general (i.e., a general measure of reading comprehension that included both inferential and literal items but did not include separate data for each type of comprehension). The inferential measures ranged in difficulty across studies and included items such as "Why did Billy cry even more?" (Yuill & Oakhill, 1988, p. 38), "Do you think new laws are needed to ensure equal rights for all Americans? Why or why not?" (Spires & Donley, 1998, p. 251). Literal measures included items such as "Where was Billy?" (Yuill & Oakhill, 1988, p. 38) and "What causes rheumatic fever?" (McNamara et al., 2006, p. 154). This review was designed to consider the impact of interventions that could be useful in understanding and learning from texts in real classrooms, so only studies using passage-level comprehension measures were selected. Studies with sentence level outcomes for comprehension were excluded. Both custom and standard measures of reading comprehension using multiple-choice questions, cloze, open-ended, and recall measures were included. We also included passage-level measures of listening comprehension, because these measures have been shown to be highly correlated with reading measures (van den Broek et al., 2005). In addition, measures had to be reported in a quantitative format that allowed calculation of an effect size using the standardized difference between means.

Participants. Studies had to be conducted with school-age students in Grades K–12. Studies with second language learners were excluded to avoid the possible confounds of differences because of learning in a second language.

Research design. Experimental and quasi-experimental designs were eligible for inclusion. Studies had to include a posttest control design with randomization or a pretest–posttest control group design. One-group pre-post designs were excluded. Studies had to use a business as usual control, with or without exposure to materials, or a weaker intervention used to mirror typical classroom practice (e.g., main idea instruction). Studies comparing skilled and less skilled readers were included if they used separate comparable control groups.

Identification and Retrieval of the Reports

A comprehensive search of the extant literature was conducted in an effort to obtain the entire population of studies focused on improving the inference making abilities of students that met the inclusion criteria. An electronic search using ERIC and PsycINFO was conducted with the search terms *inference AND comprehension, inference AND reading*, and *inference AND instruction* that yielded 2,075 citations. In addition to electronic searches, the reference lists of the included reports were searched to capture any missing articles or chapters not identified in the electronic searches. After reviewing each of the abstracts and other identified chapters for inclusion, a total of 308 articles were considered relevant. These articles and chapters were obtained, read, and further evaluated for inclusion. This resulted in 25 studies that met the eligibility criteria. Many studies were near misses including studies that lacked a comparable control (e.g., Cain et al., 2001; Carr & Thompson, 1996; Golden, Gersten, & Woodward, 1990; McKenzie, 1972), lacked enough information to compute an effect size (e.g., Dewitz et al., 1987; McGee & Johnson, 2003; McNamara et al., 2006; Oakhill & Patel, 1991; Raphael, 1984; Raphael & McKinney, 1983; Raphael & Wonnacott, 1985; Sinatra, Beck, & McKeown, 1993; Shoop, 1982), provided inadequate information to evaluate whether the components of the intervention were focused on inference instruction (e.g., Beebe & Malicky, 1982), or provided inadequate information to determine if the assessments provided information about students' comprehension (e.g., King, 1994; King et al., 1998; King & Rosenshine, 1993). Some of the studies were found in multiple reports. Many of the duplicated studies were published from dissertations (e.g., Carr, Dewitz, & Patberg, 1983; Hansen, 1981; Hansen & Pearson, 1983), or were research findings summarized in teacher journals or book chapters (e.g., Holmes, 1983; Yuill & Oakhill, 1991). If two reports were found for the same experiments, the earlier report was coded.

Coding the Research Reports

All eligible studies were coded separately for study characteristics and effect sizes so that knowledge about the results would not influence coding for the characteristics. All reports were coded by the author, another reading researcher, and doctoral student who were all trained in meta-analysis and familiar with the reading literature. Interrater agreement was determined using percentage agreement. Agreement across categories ranged from 85 to 100% with an overall average of 91%. The coders reconciled all disagreements by reviewing and discussing each discrepancy. The variable with the lowest reliability was the type of inference taught in the intervention. Some of these errors were because of initial ambiguous coding definitions or missing the information in the reports.

Effect Size Coding

Comprehension outcomes. First, each report was coded by selecting a passage-level outcome measuring the construct of inferential, or literal comprehension. Some studies reported only general comprehension effects, whereas others provided both inferential and literal effects. To retain data independence, only one outcome for each construct (i.e., general comprehension, literal comprehension, and inferential comprehension) could be considered. To retain independence, if more than one measure was reported per construct (e.g., a standardized measure that provided separate information for literal and inferential items), the measure providing separate information for literal and inference outcomes was selected over other measures. Some experiments reported scores on worksheets or passages with questions interspersed throughout the intervention, as well as, a posttest. In these cases, we selected the posttest. For cases in which the multiple measures or questions within a study were comparable, measures were aggregated using an approach that takes into account sample size differences between conditions by weighting the groups accordingly (Nouri & Greenberg in Cortina & Nouri, 2000).

Condition selection. An independent set of effect sizes was created by selecting only one condition per experiment. This selection process avoided the problem of data dependency resulting from the use of a common control. The condition with the most elements or the one considered most intensive according to the author's description was retained.

Subgroups. To better understand the differential effects because of students' reading ability, studies reporting results for skilled and less skilled readers with comparable control groups were treated as separate effects. In addition, grade-level effects were considered independently when studies reported both treatment and control groups for each grade level.

Calculating Effect Sizes

If an author reported a *d-statistic* effect size, the effect size was retained without change. In other cases, we calculated the *d-statistic* by taking the difference between the intervention group and the control group means and dividing by the pooled *SD*s of the means. Whenever possible, d was calculated after adjusting for differences at pretest (What Works Clearinghouse, 2014). In instances where *SD*s were not reported, they were estimated from reported t-statistics (see Shadish, Robinson, & Congxiao, 1999; Smith, Glass, & Miller, 1980) or residual sums of squares. One study (i.e., Yuill & Oakhill, 1988) reported means without *SD*s forcing a conservative estimate of effects using the exact p values reported from appropriate t- or F tests, and one study required estimation using the residual sums of squares (i.e., Fortenberry, 1984).

Moderator Coding

General study characteristics. Methodological characteristics were coded for each study, so differences because of study design or implementation could be ruled out as possible confounds. Coded variables included the year the study was published, whether participants were randomly assigned to conditions, and the amount of attrition that occurred. The type of control group used was coded as a condition that used the same materials or a business-as-usual control. It is likely that effects comparing inference instruction and business-as-usual controls would be greater in magnitude than control conditions in which students used similar materials.

Dependent measure. Each measure was identified as inferential, literal, or general. The measure was also coded for reliabil-

ity, the type (i.e., listening or reading), and format (i.e., recall, open-ended, and multiple-choice) of the measure. After reviewing the studies, we decided not to code for delayed effects, because of the limited number of studies measuring delayed effects and the variability in the amount of time between the end of instruction and the administration of the delayed measures. For example, some delayed measures were administered one week after instruction, while others were administered up to four months after instruction.

Participant characteristics. Past research has shown that good and poor readers differ in their ability to make inferences (Cain & Oakhill, 2007). Some studies have shown that poor readers make more gains after instruction than good readers (Yuill & Joscelyne, 1988; Yuill & Oakhill, 1988). Whereas other studies have shown that good readers make more gains on inference tasks and poor readers make more gains on literal tasks after receiving instruction in inference generation (Hansen & Pearson, 1983). Therefore, participants were coded as skilled readers or less skilled readers as indicated by the authors of the studies. Authors used the terms poor readers and poor comprehenders to refer to students who show deficits in word level reading and/or comprehension. Often authors used a general reading achievement test to establish reading ability. We also know that the ability to make appropriate inferences increases with development (Ackerman, 1988; Paris & Lindauer, 1976); therefore, grade level was coded to consider these differences. Other participant characteristics including SES (low or middle/high), gender, and ethnicity were also coded.

Type of instruction. Inference instruction was considered broadly in this review to include any method and procedure intended to enhance inference generation and promote comprehension. The instructional categories coded can be found in Table 1.

Contextual intervention factors. Contextual factors such as the total hours of instruction, who implemented the instruction (i.e., researcher, teacher, or other), and group size (i.e., one-to-one, small group of 10 or less, or whole class of 11 or more) were coded. In addition to the contextual factors, we classified and coded the type of instruction used in each study. This was difficult, because most of the included studies used multiple instructional elements. We coded all of the instructional elements reported, but we also decided to classify instruction by the element for which the researchers allocated the most time. This allowed us to assess any differential effects because of instructional methods.

Type of inference. The type of inference taught during instruction may also impact results. There is evidence that, in general, local or text-based inferences that require integration of information stated explicitly in adjacent text are easier to make than global or elaborative inferences that require using background knowledge to fill in missing conceptual gaps across a text (Bowyer-Crane & Snowling, 2005; Cain & Oakhill, 1999; Olson, 1985). Despite the extensive extant literature on the nature of inference generation, there is no agreed upon definition or system of classification (e.g., Graesser et al., 1994; Nicholas & Trabasso, 1980; Pearson & Johnson, 1978; Perfetti & Stafura, 2015). After reviewing the included studies, we noted that some studies provided enough information to classify the taught inference into one or more of the taxonomies, but others taught multiple types of inferences or did not provide enough details to identify the specific type of inference taught. Therefore, we decided to use the categories of inference accepted by most researchers: (a) text-based inferences which are required to connect textual elements to produce a coherent text-based representation, and (b) elaborative inferences that are not required to understand the text-based representation of the text, but facilitate a more complete situation model of the text.

Additional instructional elements. Reading comprehension researchers have found that explicit instruction improves reading comprehension and encourages generalization (Pearson & Dole, 1987). Explicit comprehension instruction is a model for teaching in which the teacher directly models the skill, guides the students through the acquisition of the new skill by providing decreasing levels of support as they gain proficiency, and then encourages students to internalize the strategy through practicing and applying the skill independently. Some researchers have attributed the inconsistent results of inference instruction to the lack of explicit instruction (Pearson & Dole, 1987). In addition, across theoretical stances, researchers acknowledge the importance of metacognitive skills in the ability to generate inferences. We, therefore, coded whether any metacognitive elements were incorporated into the intervention. Writing instruction has been shown to increase reading comprehension (Graham & Hebert, 2011), as well as discussion (Kucan & Beck, 1997) so these elements were also coded.

Text type. The type of text used (i.e., expository or narrative) could also explain differential effects across studies. Expository texts are considered more difficult than narrative texts, because

Table 1
Description of Inference Instruction

Instruction	Description
Inference question practice and question generation	Students practice answering or generating inferential questions during or after reading.
Text clue integration strategy	Students are taught how to use clues in the text to construct a coherent representation.
Background knowledge strategy	Students are explicitly taught to use relevant background knowledge to fill in gaps in the text.
Text structure and organization	Graphic organizers are used to make the text structure clear to students or to activate background knowledge.
Inference types	Inference is taught by breaking down inferences into subcategories and practicing each skill individually.
Self-explanation or elaboration	Students learn to connect text ideas with prior knowledge through self-explanation and elaboration.
Perspective taking	Students are taught to empathize with characters and consider character motives to make predictive and causal inferences.

they require more intense processing because of their demands on content knowledge, abstraction, and context-specific vocabulary (Graesser, Golding, & Long, 1991; Olson, 1985; Wolfe, 2005). We coded texts as narrative, expository, or both.

Implementation fidelity and training. It is important to establish that instruction was implemented as intended by trained instructors. Therefore, fidelity of intervention implementation of instruction and the training of the instructors were coded. This variable was coded liberally. If the authors mentioned monitoring instruction in any manner the study was given credit for fidelity. Likewise, if the authors mentioned training, coaching, or professional development, the study was given credit for having trained instructors.

Statistical Procedures

The analyses were separated into three outcomes of interest (i.e., general, inferential, and literal comprehension). Each distribution of effects and sample sizes was examined to determine if there were any outliers that could distort results. Outliers were identified as being 2 *SD*s from the mean. To limit the influence of these values on subsequent analyses, outliers were Winsorized as recommended by Lipsey and Wilson (2001). No outliers were identified for the general, inferential, or literal effect sizes, but one study was identified as an outlier for sample size with both inferential and literal outcomes (Elbro & Buch-Iversen, 2013; $n = 151$). We trimmed this effect by identifying the next lowest sample size ($n = 73$), and set the value at the midpoint between the two values ($n = 117$). This allowed the study to remain the largest sample size, but reduced its influence.

Effect sizes derived from small samples are known to be biased, so the effect sizes were adjusted using a small sample correction, $1 - (3/4n - 9)$, where n is the total sample size for computing each effect (Hedges, 1982). Each *Hedges' g* effect size was then weighted by the inverse of its error variance, $1/SE^2$, to take its proportionate reliability into account (Shadish & Haddock, 1994).

Next, the overall main effects for literal, inferential, and general outcomes were estimated using a random effects model. A test of homogeneity using the Q-statistic was then applied separately for each outcome to establish whether there was more variability in the effects than would be expected by subject-level sampling error alone (Hedges & Olkin, 1985). The Q-statistic is calculated as

$$Q = \sum w_i \times (ES_i - \mu_{..} ES)^2$$

in which w_i is the inverse variance weight for each effect size i, and ES_i is the weighted mean effect size for each i, and ES is the weighted mean effect size over all cases of i. This statistic is distributed as chi-square with $k - 1$ degrees of freedom where k is the number of effect sizes and when significant, warrants rejection of the null hypothesis that variance in effects are explained by sampling error alone (Lipsey & Wilson, 2001). For outcomes whose variance exceeded that predicted by sampling error alone (i.e., Q, $p < .05$), mixed-weight regression analyses were conducted to estimate the moderating influence of method, participant, and intervention variables on literal and inferential comprehension outcomes (Raudenbush, 1994).

Results

Descriptive Characteristics of Studies

General study and participant characteristics. The descriptive information for general study and methodological characteristics, participant, and intervention characteristics across studies are shown in Appendix A. It should be noted that not all of the variables of interest could be included in the moderator analysis because of the limited number of studies found from the literature search, but descriptive information about these study characteristics are included to provide as much information about the inference studies as possible.

The literature search yielded 25 eligible reports from which 25 inferential, 18 literal, and 13 general effects were derived. The majority of the studies were conducted in the 1980s in peer-reviewed journals. Most studies were short, with almost half lasting fewer than 5 hr and more than 70% lasting 10 or fewer hours. The majority of the interventions were conducted by researchers in a classroom setting. Across these studies, there were 1,752 participants. Most of the participants were skilled readers from middle SES backgrounds in the 3rd to 8th grade. No inference studies were found for kindergarten or Grades 1, 10, 11, and 12. Most articles did not report gender or ethnicity.

Methodological characteristics. Most of the studies used a randomized design with control groups that had exposure to the materials the experimental groups used. The majority of the outcome measures was experimenter-designed and used an open-ended question or a recall format. Only a few effects ($k = 7$) across five studies were derived from norm-referenced standardized measures. The majority (more than 75%) of the studies reported reliability information for measures. Most studies using custom measures assessed reliability using interrater reliability and internal consistency (i.e., Cronbach's α, KR-20). In addition, although some type of reliability information was reported for the majority of the studies, almost two thirds failed to provide evidence of training or coaching for the personnel implementing the intervention or for monitoring implementation fidelity.

Intervention characteristics. Many of the methods used in the reviewed studies aligned with the theoretical and cognitive literature concerning inference generation. Many of the studies focused on teaching students to locate relevant information in text to generate an inference, to integrate information across text, to provide evidence in the text of their answers to inferential questions, or to activate and integrate background knowledge with information in the text. A majority of studies combined several instructional features (e.g., metacognitive strategy instruction and background knowledge activation) to teach inference generation. Nine of the studies in this review primarily used text clue instruction in which students were taught to find the most relevant information in the text. Some methods used a cloze procedure in which students were required to provide a missing word in the text by considering the words around it. In other text clue instruction, students looked for clue words to figure out what was happening in the passage. Other interventions focused on activating or supporting background knowledge using advanced organizers or structural overviews to activate children's background knowledge. In addition, some studies focused on improving children's awareness and use of connecting prior knowledge to information in the

text. While most of the studies included practice in answering inferential questions, some asked students to generate questions. Only a few studies assessed the effects of self-elaboration or instruction on different types of inferences. For detailed information about effect sizes, participants, and instructional elements per study, see Table 2.

Overall Effects

The general comprehension effect sizes (k = 13) ranged from −0.28 to 1.77 for an overall random weighted mean effect size of 0.58, which was significantly different from 0, $p < .01$ (see Appendix Figure C1). The inferential comprehension effect sizes (k = 25) ranged from −0.18 to 2.24 with an overall random weighted mean effect size of 0.68, which was significantly different from 0, $p < .01$ (see Appendix Figure C2). The literal effect sizes (k = 18) ranged from −0.46 to 1.85 with an overall random weighted mean effect size of 0.28, which was significantly different than 0, $p = .04$ (see Appendix Figure C3).

Moderator Effects

General comprehension. Moderator analyses were conducted to determine if there were systematic differences among studies that could be identified. First, a Q-test was conducted for the general comprehension effects. The test was not significant, Q (12) = 17.90, $p = .12$, indicating that there was no heterogeneity in the effects that could be attributed to the effects beyond sampling error. Therefore, no moderator analyses were conducted with the general comprehension outcomes. Although no moderator analyses were conducted for the general outcomes, a separate analysis to consider the overall effect size for standardized measures using a subset of the general effects was conducted. The overall effect for standardized measures was $d = 0.53$.

Inferential comprehension. For inferential outcomes, the Q-test was significant, Q (24) = 69.50, $p < .01$, indicating heterogeneity in the effects and warranting a moderator analysis. It should be noted that many of the methodological, measurement, participant, and instruction characteristics were not reported or there were too few ($k < 5$) to use in subsequent moderator analyses (e.g., attrition, measure reliability, and standardized measure).

First, the methodological characteristics (i.e., control group strength, random assignment, pre-post effect size adjustment, and publication year) were examined to determine if any were associated with effect size and would need to be controlled in subsequent analyses. Zero-order random effects correlations of method variables and effect size were conducted using method of moments estimation (see Table 3). Although these correlations are helpful in considering

Table 2
Methodological and Intervention Characteristics by Study

	Literal effect size	Inference effect size	General effect size	Standardized measure*	Treatment N	Control N	Control group strength	Average grade	Type of reader	Treatment hours	Type of inference	Type of text	Inference question practice
Bailey, Silvern, Brabham, & Ross (2004)		1.59			26	31	BAU	2		15	E/T		
Carnine, Kameenui, & Woolfson (1982) a		1.29			12	12	BAU	5	PC	3	E/T	NAR	
Carnine, Stevens, Clements, & Kameenui (1982) b		2.24			10	10	BAU	5	PC	2	E	NAR	
Carr (1983) Experiment 1	.06	.20			24	21	BAU/M	6		27	E/T	EXP	★
Carr (1983) Experiment 2	.23	.56	.39*	ITBS	22	17	BAU/M	6		27	T	EXP	★
Davey and McBride (1986)	−.08	.10			23	24	BAU/M	6		3	T	EXP	★
Dyck and Sundbye (1988)	1.45	.96			12	12	BAU/M	6	LD	2	E/T	NAR	★
Elbro and Buch-Iversen (2013)	.44	.72			151	85	BAU	6		8	E	EXP	+
Emery and Milhalevich (1992) 4th grade			−.28		21	21	BAU/M	4		4	E	NAR	+
Emery and Milhalevich (1992) 5th grade			.20		23	23	BAU/M	5		4	E	NAR	+
Emery and Milhalevich (1992) 6th grade			1.38		23	23	BAU/M	6		4	E	NAR	+
Fortenberry (1984)	−.17*	.21*		SDRT	55	63	BAU	8		15	E	NAR	
Gordon (1980)	.42	.56	.20*	MAT	14	14	BAU/M	5		20	E	NAR	
Hansen (1980)	.29	.55	.63*	SAT	8	8	BAU/M	2		6	E		+
Hansen and Pearson (1983) less skilled	.70	.99			10	10	BAU/M	4	LS	7	E	both	
Hansen and Pearson (1983) skilled	−.28	1.07			10	10	BAU/M	4		7	E	both	
Holmes (1985)		2.01	.83*	NRT	6	6	BAU/M	4.5	LD	3	E	EXP	+
McCormick and Hill (1984)	.35	−.18	.19*	MAT	23	23	LT	5	LD	67	E	both	+
McNamara et al. (2006)			.41		17	21	BAU	8		2	T/E	EXP	
Price-Cheves (1973)		.06*		NDRT	47	45	BAU/M	6		10	E		+
Reutzel and Hollingsworth (1988) less skilled	1.85	1.16			25	25	BAU	3	LS	16	T	NAR	
Reutzel and Hollingsworth (1988) skilled	−.29	.41			25	25	BAU	3		16	T	NAR	
Seifert (1993)	−.46	.73			29	28	LT	7		2	E	EXP	+
Spires and Donley (1998) Experiment 2	−.20	.48			35	39	BAU/M	9		5	E	EXP	
Spires and Donley (1998) Experiment 1	−.06	.43			24	30	BAU/M	9		5	E	EXP	
Stitt (1968)		.98			73	73	BAU	6		8	E	EXP	
Sundbye (1987)	.97	1.22			12	12	BAU/M	3		2	E	NAR	★
Wells (1986)		.21			67	44	BAU/M	8		7	E	both	★
Winne, Graham, and Prock (1993)	.64	.65			11	11	LT	4	LD	9	T	both	
Yuill and Oakhill (1988) less skilled			1.77		9	9	LT	2	PC	4	T	NAR	
Yuill and Oakhill (1988) skilled			.50		10	10	LT	2		4	T	NAR	
Yuill and Joscelyne (1988) less skilled Experiment 2			.92		9	9	BAU/M	2	PC	4	T	NAR	
Yuill and Joscelyne (1988) skilled Experiment 2			.60		10	10	BAU/M	2		4	T	NAR	

Note. ITBS = Iowa Test of Basic Skills; SDRT = Stanford Diagnostic Reading Test; MAT = Metropolitan Achievement Test; SAT = Stanford Achievement Test; NRT = Nelson Reading Test; BAU = Business as Usual Control; BAU/M = Business as Usual Control with Materials; LT = Lesser Treatment; PC = Poor Comprehender; LD = Learning Disabled; LS = Less Skilled Reader; E = Elaborative; T = Text Based; NAR = Narrative; EXP = Expository.
* Indicates an effect size derived from a norm-referenced standardized test. ★dominant instructional component. +instructional component present.

Table 3
Standardized Correlation Coefficients of Selected Method Variables for Inferential (K = 25) and Literal (k = 17) Outcomes

Method variable	$\beta_{inference}$	$\beta_{literal}$
Control group strength	.39	.15
Random assignment	.09	.19
Pre-post effect size adjustment	−.05	−.14
Publication year	.12	−.04

Note. Random-weighted analysis; β [μτ] .15 were considered non-trivial and were carried forward in subsequent analyses. Analysis included control group strength (0 = business-as-usual control; 1 = same materials); random assignment (0 = quasi-experiment; 1 = random assignment or matched design); pre-post effect size adjustment (no pre-post adjustment = 0; pre-post = 1).

which method variables might be influencing effects, caution must be exercised in their interpretation, because they ignore all other moderating factors. Only one methodological variable, control group strength, showed a nontrivial relationship (i.e., $\beta > .15$; Lipsey & Wilson, 2001) and was carried forward as a control moderator.

Next, a series of inverse-variance weighted random effects multiple regressions were used. To identify relationships between effect size and study characteristics without the confounding influence of other study characteristics (see Raudenbush, 1994), each characteristic was considered separately while controlling for the method moderators that were carried forward (see Table 4). For the inferential effects, in addition to the control group strength, only three other moderators were associated with effect size. Grade level was negatively associated with effect size indicating that inference interventions with younger students were more effective than interventions with older students. The moderator analysis also showed that less skilled readers benefited more from inference instruction than skilled readers, and that instruction in small groups was more effective than instruction in larger formats.

Next, the significant moderators (i.e., control group strength, grade level, reading ability, and group format) were simultaneously entered to take into account relationships among variables (see Table 5). In this final model, only control group strength and small group instruction were significantly associated with inferential effects.

Literal comprehension. A Q-test was conducted for the literal effects and was found to be significant, $Q\ (17) = 61.42$, $p < .01$, indicating heterogeneity in the effects that could be identified in a moderator analysis. Using the same procedures as above, we considered the effects of moderators on the literal comprehension outcomes.

First, we examined the methodological characteristics (i.e., control group strength, random assignment, pre-post effect size ad-

Question generation	Facilitated questioning	Text clue integration	Text structure or organization	Background knowledge strategy	Inference type instruction	Strategy instruction	Self explanation	Metacognitive monitoring	Perspective taking	Direct explanation and modeling	Discussion	Writing
					★						+	
	★					+						
	★					+				+		
		+	+			+		+		+	+	
		+	+			+		+		+	+	
+											+	
				★	+				+		+	
									★		+	+
									★		+	+
									★		+	+
★												+
	+			★	+		+				+	
				★		+		+				
				★		+		+		+	+	+
				★		+		+		+	+	+
+		★				+	+	+		+		+
				★		+				+		
						+	★	+		+		
		★									+	
		★			+		+	+	+	+	+	+
		★			+		+	+	+	+	+	+
				★								
				★		+			+	+	+	
				★		+			+	+	+	
					★						+	
+												
+		+						+				
		★				+					+	+
+		★				+						
+		★				+						
		★										
		★										

Table 4
Conditional Correlations for Inferential (k = 25) and Literal (k = 18) Comprehension Outcomes Controlling for Method Variables

Study characteristic	$\beta_{inference}$	p	$\beta_{literal}$	p
Participant				
Grade level	−.42	.02	−.42	.07
Less skilled vs. skilled	.40	.03	.78	<.01
SES	−.11	.59	−.18	.49
Instruction, measure, and text				
Small vs. large group	.66	<.01	.51	.03
Instructional intensity (hours)	−.33	.08	.13	.62
Activating background knowledge	.00	.99	−.37	.17
Text clues	.13	.50	.25	.35
Questioning	−.01	.99	.23	.36
Meta-cognitive strategy	−.02	.93	.16	.54
Elaborative vs. text-based inference	.11	.50	−.13	.61
Explicit instruction	.02	.94	−.10	.70
Discussion	.04	.82	.13	.60
Writing	−.02	.93	.04	.89
Listening measure	.32	.08	—	—
Open-ended questions vs. other	.26	.22	—	—
Expository text vs. other	−.20	.17	−.42	.07

Note. Mixed-weighted analysis. Each instructional characteristic was entered individually and tested in the presence of method moderators (for inference effects: control group strength; for literal effects: control group strength and random assignment). SES = socioeconomic status; analysis included: SES (high/middle vs. low); small group (one-to-one or small group with 10 or less students) vs. large group (11 or more students); each instructional type or element vs. other. Listening measures for literal outcomes or open-ended questions could not be considered because of a limited number of studies ($k < 5$).

justment, and publication year) to determine if any were associated with effect size and would need to be controlled in subsequent analyses. Zero-order random effects correlations of method variables and effect size were conducted using method of moments estimation (see Table 3). For the literal effects, control group strength and random assignment were practically associated with effect size (i.e., $\beta > .10$) and carried forward as control variables in subsequent analyses.

Next, a series of inverse-variance weighted random effects multiple regressions were conducted. Each characteristic was considered separately while controlling for the method moderators that were carried forward (see Table 4). In contrast to the inferential effects, grade level approached but did not reach significance. Similar to the inferential outcomes, less skilled readers benefited more from inferential instruction than more skilled readers on literal outcomes, and small group instruction was more effective than whole group instruction. The active moderators for literal comprehension were then considered simultaneously in a final model (see Table 6). In this model, only control group strength and reading ability explained differences in literal effects.

Comparison of skilled and less skilled readers. To make clearer the comparison of skilled and less skilled readers, an overall mixed weighted mean effect size for each type of reader for inference and literal effects was computed. Without considering conditionality of method moderators, the overall effect for skilled readers on inference measures was moderate, $d = 0.55$, and the effect for less skilled readers was large, $d = 0.80$ (Cohen, 1988; Appendix Figure C2). The overall effect for skilled readers on literal measures was not significantly different than 0, $d = 0.06$, but the effect for less skilled readers was large, $d = 0.97$ and significantly different than 0 (see Appendix Figure C3). Although skilled and less skilled readers benefited similarly from inference instruction, only less skilled readers benefited substantially on literal measures of comprehension after receiving inference instruction (see Figure 1).

Publication Bias

One threat to the conclusions of any meta-analysis that must be addressed is publication bias (e.g., Sterne & Egger, 2001; Sutton, 2009). The mean effect estimates may be upwardly biased because of nonreporting of underpowered studies or studies producing negative effects. Publication bias can be identified through visual analysis of the effects plotted by their *SE*. We expect the plotted effects to be symmetrical around the mean and form an inverted funnel. Visual analysis of the funnel plots (see Appendix B) for the general and inference outcomes showed a few outlying effects, as well as a number of large effects not predicted by their *SE* that suggests possible publication bias. However, the effects for the literal outcomes were more evenly distributed and show little indication of publication bias. Likewise, Egger's regression intercept for the general effects was 2.90 ($p = .04$, one-tailed), for the inferential effects was 1.79 ($p = .03$, one-tailed), and for the literal effects was 1.42 ($p = .14$, one-tailed) indicating no bias. To consider the extent to which the overall effect sizes for the general and comprehension outcomes are potentially impacted by publication bias, Duval and Tweedie's trim and fill procedure (Duval & Tweedie, 2000) was conducted. This method determines the number of studies needed to be trimmed from

Table 5
Final Model for Inferential Comprehension Outcomes (k = 25)

Study characteristic	B	*SE*	95% CI Lower limit	95% CI Upper limit	β	p
Method						
Control group strength	.34	.15	.04	.64	.37	.05
Participant						
Grade level	−.01	.32	−.12	.11	−.10	.67
Less skilled vs. skilled	.10	.26	−.42	.61	.09	.64
Instruction						
Small vs. large group	.36	.15	.06	.66	.56	.01

Note. Mixed-weighted analysis. $Qresid(20) = 11.70$, $p = .93$.

Table 6
Final Model for Literal Comprehension Outcomes (k = 18)

Study characteristic	B	*SE*	95% CI Lower limit	95% CI Upper limit	β	p
Method						
Control group strength	.37	.18	.01	.72	.42	.04
Random assignment	.94	.30	−.52	.58	.02	.92
Participant						
Less skilled vs. skilled	.94	.30	.34	1.54	.68	<.01
Instruction						
Small vs. large group	.24	.18	−.11	.60	.19	.28

Note. Mixed-weighted analysis. $Qresid(13) = 8.93$, $p = .77$.

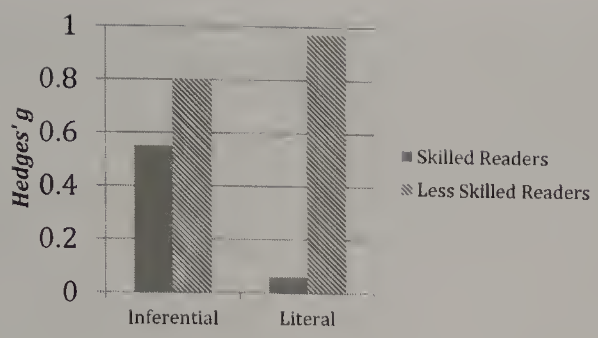


Figure 1. Comparison of inferential and literal effects for skilled and less skilled readers.

the right side of the plot and imputed on the left to produce symmetry. The trimmed studies are imputed as mirror images on the left side of the plot. Using this procedure with the general outcomes, five studies were imputed resulting in an estimated effect size of reduction from $d = 0.58$ to $d = 0.25$. For the inference outcomes, six studies were trimmed and imputed resulting in a reduction of the overall effect size from $d = 0.68$ to $d = 0.48$.

Publication bias cannot be determined definitively, and in some cases, the effects may not be caused by publication bias, but may represent a true shift in the effects due the interventions conducted in the smaller studies (Borenstein, Hedges, Higgins, & Rothstein, 2011; Sutton, 2009). In this review, studies with fewer participants tended to use smaller groups, be conducted with less skilled readers, and use more intense types of instruction. Other differences associated with these studies may not have been captured in the moderator analysis and may account for the asymmetry of the plots for the general and inference. In addition, the trim and fill sensitivity analysis showed that if publication bias was present the point estimates would be positive for general and inference outcomes.

Discussion

Overall, the effects of inference instruction on general and inferential comprehension outcomes were moderate to large for skilled and less skilled readers. These findings align with previous research supporting the view that inference generation plays a causal role in reading comprehension. The findings also revealed that inference instruction implemented with less skilled readers improved not only inferential understanding, but also literal comprehension of text. Although higher order skills are considered difficult to teach, surprisingly, most of the studies showed positive results in relatively short periods of time (i.e., less than 10 hr). One instructional factor, providing instruction in a small group, stood out as being important to inference outcomes.

Theory and Intervention Development

Inference generation requires the coordination of many cognitive processes (e.g., background knowledge activation, integration processes, suppression of irrelevant information, working memory, and executive function skills). Many of these aspects were targeted for instruction in the reviewed studies. Inference generation has been shown to be highly dependent on background knowledge (Kendeou et al., 2014; Kintsch & Kintsch, 2005) and integration processes (Barnes et al., 1996; Cain et al., 2001). More than two-thirds of the reviewed studies included background knowledge strategies or text clue integration procedures as primary components of the interventions, reflecting the importance of these skills to inference generation. The importance of executive function skills was also acknowledged across the studies. Although not a primary component in most studies, the majority of the interventions included procedures to explicitly teach students to use strategies to make inferences, and in almost half of the studies, children were explicitly taught to monitor their understanding as they read.

Two of the cognitive skills addressed in some models of comprehension, working memory and suppression, were not mentioned in the majority of studies. Although they were not explicitly addressed, some of the studies used procedures to support these skills. For instance, most of the studies used procedures likely to support students with working memory deficits such as partitioning the text into smaller sections and asking questions after each section instead of waiting until the end of an entire passage. An interesting find was that although struggling readers have been known to make haphazard elaborations (Williams, 1993) and have difficulty suppressing irrelevant knowledge in making inferences (Gernsbacher & Faust, 1991), only a few studies mentioned any procedures to provide feedback to help students suppress irrelevant knowledge. Reutzel and Hollingsworth (1988) required students to justify the inferences they made by identifying evidence in the text, and McNamara et al. (2006) emphasized the importance of closely linking what students know to information in text. On the other hand, to encourage knowledge activation, Spires and Donley (1998) purposefully made no initial judgments about the knowledge the students generated. They did, however, ask students who related off topic prior knowledge, "Can you explain why that portion of text reminded you of that specific experience or information?" (p. 251). Given the problems with struggling readers haphazard elaborations and variability in the procedures used to help students select and use relevant prior knowledge, future studies should attempt to identify the most promising procedures for helping students suppress irrelevant knowledge.

Overall this review showed that inference interventions are effective at increasing general and inferential comprehension. Unfortunately, we were unable to discern which interventions were most promising. Many of the studies contained multiple components intended to increase inference generation. This approach makes sense from a practical standpoint, as multiple strategies have a better chance of addressing the various cognitive skills required to successfully generate inferences. However, the overlap of these instructional components muddies any comparisons of specific inference strategies and is likely the reason why no particular type of instruction was found to be more effective than any other in this review. Future intervention research should consider isolating specific instructional procedures using a component analysis to determine which methods are most beneficial.

Differential Effects for Skilled and Less Skilled Readers

Although both skilled and less skilled readers benefited from inference instruction on general comprehension and inferential outcomes, there were marked differences between how they responded to literal questions. Keeping in mind that the studies selected for this review focused on strategies and methods for increasing inferential understanding, it was somewhat unexpected that less skilled readers showed substantial growth on an

aspect of comprehension that was not explicitly taught. Conversely, this finding makes sense when considering the problems less skilled readers exhibit when engaging with text. These readers are often characterized as more passive than skilled readers (e.g., Cain & Oakhill, 1999). Many of the studies in this review provided explicit instruction in finding pertinent information in a text and integrating it with prior knowledge to answer inferential questions. It seems that this type of instruction would be especially beneficial to less skilled readers, as it would aid them in making inferences and require them to attend to important details to which they may not have otherwise noticed. Generating inferences also requires students to engage with the text at a deeper level than answering factual questions. In memory research, deeper levels of processing have consistently been shown to lead to higher levels of retention and greater retrieval (Craik, 2002; Craik & Lockhart, 1972). This finding is consistent with Kintsch's CI model (1988) of comprehension in which the generation of inferences can strengthen memory traces for the literal content.

In contrast, the skilled readers made no gains in literal understanding. It may be that at pretest, the skilled readers were able to form a textbase level understanding of what they read and were therefore competent at answering the literal questions. If readers were already competent, they would be unlikely to show growth in answering literal questions after receiving instruction in inferencing. Alternatively, teaching students who generally have an adequate understanding of the textbase to focus exclusively on processing at the situation model may encourage less attention to be paid to the details in the text and may weaken their performance on literal questions. Future experimental studies informed by discourse processing models and memory paradigms should be conducted to consider the possible impact of inference instruction on readers' literal understanding of text.

Limitations of the Review

The ability to make inferences is considered the hallmark of a good reader, and a vast amount of research has been conducted to better understand inference generation. A surprising finding from this review was that so few well-designed intervention studies have been conducted to address this important skill. As discussed earlier, one reason for the limited number of studies may be because of publication bias and lack of reporting by researchers. Many inference studies were excluded, because there was not enough information available to compute an effect size. It should be noted that the majority of the excluded studies reported positive effects that are consistent with the overall findings of this review. However, the inclusion of these studies might have changed some of the results in the moderator analysis. Another limitation is that the results of this review can only be generalized to students in Grades 2 through 9. There were no studies conducted in the earlier or later grade levels. Studies will have to be conducted in the primary grades and high school level to assess if inference instruction is beneficial for these students.

Another limitation of this review concerns not taking into account whether students, especially those with reading difficulties could adequately access the texts used in the studies. High quality lexical representations have been shown to support inferential processes and comprehension (Perfetti & Stafura, 2014; García & Cain, 2014). Some of the studies controlled for word-level processes through reading the passages aloud. Others report the characteristics of the texts including grade level, but most of the studies did not have enough information to reliably code whether readers could easily read the text. It should also be noted that the text chosen for studies conducted with less skilled readers may have been easier than texts used in other studies. Using easier texts may have made inferential processing easier and inflated results for those studies. A related limitation of the review concerns issues with lumping different types of readers (e.g., struggling readers, poor comprehenders) together in the analysis. This was done for two reasons. First, there were not enough studies conducted with any one type of reader to analyze them separately. Second, most of the studies did not offer adequate information for classifying readers in any meaningful way.

Finally, in addition to the limited number of studies and lack of reporting, many questions could not be adequately answered because of the research methods used. Unfortunately, very few studies utilized a standardized measure of reading comprehension at posttest, and many of the custom measures were closely aligned with the instruction provided in the studies. The use of custom measures was understandable, because most of the studies were very short in duration. However, positive outcomes on standardized tests would bolster our confidence in the generalizability of effects. Likewise, the inclusion of more studies using delayed measures would also help us determine the durability of the effects.

Recommendations for Practice and Future Research

Current practices regarding the teaching of inference skills in the classroom are unknown. It is unclear if teachers are aware of evidence-based strategies for teaching inference generation and whether skilled and less skilled readers have equal opportunities to practice these skills in the classroom. The common core state standards (CCSS) emphasizes the importance of students engaging deeply with complex text. In this study, practices such as those recommended by CCSS (e.g., a focus on text dependent questions and providing evidence) were found to be effective at increasing students' comprehension. However, other interventions that taught students how to activate and integrate their background knowledge with the text were also effective. According to our current understanding of how people make sense of text, readers must use and integrate background knowledge to create an appropriate situation model. Students who have difficulty, either because they lack background knowledge or because they do not know how to integrate their knowledge with relevant information in the text, will require support if they are to be successful at independently reading complex texts. Future research should consider how to best support knowledge development and explicit instruction in inference generation. One promising approach is to design interventions that simultaneously build deep background knowledge in a topic while teaching inference skills (e.g., Barth & Elleman, 2016; McNamara et al., 2006).

Findings from this review suggest that inference skills can be effectively taught to both skilled and less skilled readers, and that for less skilled readers, inference instruction provides the additional benefit of enhanced memory for the literal content in a text. Typical instruction for teaching less skilled readers often focuses on ensuring students have an accurate representation of the textbase before asking them to answer inferential questions. Yet, findings from this review suggest that if we teach less skilled readers strategies for generating inferences, they may engage more deeply in the text and thus develop more accurate and durable textbase representations.

Comprehension instruction should be grounded in applying the discoveries made in cognitive science to instructional methodologies (Compton, Miller, Elleman, & Steacy, 2014). Future work should consider the extent to which background knowledge, integration, and suppression play a role in comprehension outcomes, and how these findings from cognitive science can lead to the next generation of comprehension methods. It would be valuable for practitioners to know which strategies hold the most promise and how these strategies are likely to interact with learner characteristics. The use of scientifically proven strategies based on our understanding of cognition will be crucial as we raise expectations for students, especially those that are struggling. Although future research will have to consider which inference generation interventions are most effective, the instructional methods reviewed in this article provide evidence that inference strategies are effective and should be utilized in schools to enhance comprehension.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Ackerman, B. P. (1988). Reason inferences in the story comprehension of children and adults. *Child Development, 59,* 1426–1442. http://dx.doi.org/10.2307/1130504

Ahmed, Y., Francis, D. J., York, M., Fletcher, J. M., Barnes, M., & Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology, 44–45,* 68–82. http://dx.doi.org/10.1016/j.cedpsych.2016.02.002

Albrecht, J. E., & O'Brien, E. J. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1061–1070. http://dx.doi.org/10.1037/0278-7393.19.5.1061

*Bailey, L. B., Silvern, S. B., Brabham, E., & Ross, M. (2004). The effects of interactive reading homework and parent involvement on children's inference responses. *Early Childhood Education Journal, 32,* 173–178. http://dx.doi.org/10.1023/B:ECEJ.0000048969.91442.36

Barnes, M. A., Dennis, M., & Haefele-Kalvaitis, J. (1996). The effects of knowledge availability and knowledge accessibility on coherence and elaborative inferencing in children from six to fifteen years of age. *Journal of Experimental Child Psychology, 61,* 216–241. http://dx.doi.org/10.1006/jecp.1996.0015

Barth, A. E., Barnes, M., Francis, D. J., Vaughn, S., & York, M. (2015). Inferential processing among adequate and struggling adolescent comprehenders and relations to reading comprehension. *Reading and Writing, 28,* 587–609. http://dx.doi.org/10.1007/s11145-014-9540-1

Barth, A. E., & Elleman, A. (2016). Evaluating the impact of a multistrategy inference intervention for middle grade struggling readers. *Language, Speech, and Hearing Services in Schools*. Advance online publication. http://dx.doi.org/10.1044/2016_LSHSS-16-0041

Beebe, M., & Malicky, G. (1982). An exploratory study into determinants of successful reading remediation. *Alberta Journal of Educational Research, 28,* 163–174.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. West Sussex, UK: Wiley.

Bowyer-Crane, C., & Snowling, M. J. (2005). Assessing children's inference generation: What do tests of reading comprehension measure? *British Journal of Educational Psychology, 75,* 189–201. http://dx.doi.org/10.1348/000709904X22674

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195154054.001.0001

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology, 2,* 331–350. http://dx.doi.org/10.1016/0010-0285(71)90019-3

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing, 11,* 489–503. http://dx.doi.org/10.1023/A:1008084120205

Cain, K., & Oakhill, J. V. (2007). Reading comprehension difficulties: Correlates, causes, and consequences. In K. Cain & J. Oakhill (Eds.), *Children's comprehension problems in oral and written language: A cognitive perspective* (pp. 41–76). New York, NY: Guilford Press.

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition, 29,* 850–859. http://dx.doi.org/10.3758/BF03196414

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96,* 31–42. http://dx.doi.org/10.1037/0022-0663.96.1.31

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology, 96,* 671–681. http://dx.doi.org/10.1037/0022-0663.96.4.671

*Carnine, D. W., Kameenui, E. J., & Woolfson, N. (1982). Training of textual dimensions related to text-based inferences. *Journal of Reading Behavior, 14,* 335–340. http://dx.doi.org/10.1080/10862968209547459

*Carnine, D., Stevens, C., Clements, J., & Kameenui, E. J. (1982). Effects of facilitative questions and practice on intermediate students' understanding of character motives. *Journal of Reading Behavior, 14,* 179–190. http://dx.doi.org/10.1080/10862968209547445

*Carr, E. M. (1983). *The effect of inferential and metacognitive instruction on children's comprehension of expository text* (Unpublished doctoral dissertation). University of Toledo, Toledo, OH.

Carr, E., Dewitz, P., & Patberg, J. (1983). The effect of inference training in children's comprehension. *Journal of Reading Behavior, 15,* 1–18. http://dx.doi.org/10.1080/10862968309547486

Carr, S. C., & Thompson, B. (1996). The effects of prior knowledge and schema activation strategies on the inferential reading comprehension of children with and without learning disabilities. *Learning Disability Quarterly, 19,* 48–61. http://dx.doi.org/10.2307/1511053

Chiesi, H. L., Spilich, G. J., & Voss, J. F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18,* 257–273. http://dx.doi.org/10.1016/S0022-5371(79)90146-4

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Compton, D. L., Miller, A., Elleman, A. M., & Steacy, L. M. (2014). Have we forsaken reading theory in the name of "quick fix" interventions for children with reading disabilities? *Scientific Studies of Reading, 18,* 55–73. http://dx.doi.org/10.1080/10888438.2013.836200

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs* (Vol. 129). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412984010

Craik, F. I. (2002). Levels of processing: Past, present. and future? *Memory, 10,* 305–318. http://dx.doi.org/10.1080/09658210244000135

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671–684. http://dx.doi.org/10.1016/S0022-5371(72)80001-X

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 99,* 311–325. http://dx.doi.org/10.1037/0022-0663.99.2.311

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Reading comprehension of scientific text: A domain-specific test of the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 102,* 687–700. http://dx.doi.org/10.1037/a0019452

Daneman, M., & Carpenter, P. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 561–584. http://dx.doi.org/10.1037/0278-7393.9.4.561

*Davey, B., & McBride, S. (1986). Effects of question-generation training on reading comprehension. *Journal of Educational Psychology, 78,* 256–262. http://dx.doi.org/10.1037/0022-0663.78.4.256

De Beni, R., Palladino, P., Pazzaglia, F., & Cornoldi, C. (1998). Increases in intrusion errors and working memory deficit of poor comprehenders. *The Quarterly Journal of Experimental Psychology: Section A, 51,* 305–320. http://dx.doi.org/10.1080/713755761

Dewitz, P., Carr, E. M., & Patberg, J. P. (1987). Effects of inference training on comprehension and comprehension monitoring. *Reading Research Quarterly, 22,* 99–120. http://dx.doi.org/10.2307/747723

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x

*Dyck, N., & Sundbye, N. (1988). The effects of text explicitness on story understanding and recall by learning disabled children. *Learning Disabilities Research, 3,* 68–77.

*Elbro, C., & Buch-Iversen, I. (2013). Activation of background knowledge for inference making: Effects on reading comprehension. *Scientific Studies of Reading, 17,* 435–452. http://dx.doi.org/10.1080/10888438.2013.774005

*Emery, D. W., & Milhalevich, C. (1992). Directed discussion of character perspectives. *Reading Research and Instruction, 31,* 51–59. http://dx.doi.org/10.1080/19388079209558095

*Fortenberry, B. H. (1984). *The use of an elaboration strategy combined with classroom television production for increasing the literal and inferential reading comprehension of eighth-grade students* (Unpublished doctoral dissertation). University of Alabama, Tuscaloosa, AL.

Franks, B. A., Mulhern, S. L., & Schillinger, S. M. (1997). Reasoning in a reading context: Deductive inferences in basal reading series. *Reading and Writing, 9,* 285–312. http://dx.doi.org/10.1023/A:1007951513772

García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research, 84,* 74–111. http://dx.doi.org/10.3102/0034654313499616

Garnham, A., & Oakhill, J. (1992). Discourse processing and text representation from a "mental models" perspective. *Language and Cognitive Processes, 7,* 193–204. http://dx.doi.org/10.1080/01690969208409384

Gernsbacher, M. A., & Faust, M. E. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 245–262. http://dx.doi.org/10.1037/0278-7393.17.2.245

Golden, N., Gersten, R., & Woodward, J. (1990). Effectiveness of guided practice during remedial reading instruction: An application of computer-managed instruction. *The Elementary School Journal, 90,* 291–304. http://dx.doi.org/10.1086/461619

*Gordon, C. J. (1980). *The effects of instruction in metacomprehension and inferencing on children's comprehension* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Graesser, A., Golding, J. M., & Long, D. L. (1991). Narrative representation and comprehension. In D. Pearson, M. Kamil, R. Barr, & P. Rosenthal (Eds.), *Handbook of reading research* (Vol. 2, pp. 171–205). Hillsdale, NJ: Erlbaum.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101,* 371–395. http://dx.doi.org/10.1037/0033-295X.101.3.371

Graham, S., & Hebert, M. (2011). Writing to read: A meta-analysis of the impact of writing and writing instruction on reading. *Harvard Educational Review, 81,* 710–744. http://dx.doi.org/10.17763/haer.81.4.t2k0m13756113566

Guszak, F. G. (1967). Teacher questioning and reading. *The Reading Teacher, 21,* 227–234.

Hannon, B., & Daneman, M. (1998). Facilitating knowledge-based inferences in less-skilled readers. *Contemporary Educational Psychology, 23,* 149–172. http://dx.doi.org/10.1006/ceps.1997.0968

*Hansen, J. A. (1980). *The effects of two intervention techniques on the inferential ability of second grade readers* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Hansen, J. (1981). The effects of inference training and practice on young children's reading comprehension. *Reading Research Quarterly, 16,* 391–417. http://dx.doi.org/10.2307/747409

*Hansen, J., & Pearson, D. P. (1983). An instructional study: Improving the inferential comprehension of good and poor fourth-grade readers. *Journal of Educational Psychology, 75,* 821–829. http://dx.doi.org/10.1037/0022-0663.75.6.821

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499. http://dx.doi.org/10.1037/0033-2909.92.2.490

Hedges, L. V., & Olkin, D. (1985). *Statistical methods for meta-analyses.* San Diego, CA: Academic Press.

Holmes, B. C. (1983). A confirmation strategy for improving poor readers' ability to answer inferential questions. *The Reading Teacher, 37,* 144–148.

*Holmes, B. C. (1985). The effects of a strategy and sequenced materials on the inferential comprehension of disabled readers. *Journal of Learning Disabilities, 18,* 542–546. http://dx.doi.org/10.1177/002221948501800909

Kendeou, P. (2015). A general inference skill. In E. J. O'Brien, A. E. Cook, & R. F. Lorch (Eds.), *Inferences during reading* (pp. 160–181). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781107279186.009

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading comprehension core components and processes. *Policy Insights from the Behavioral and Brain Sciences, 3,* 62–69. http://dx.doi.org/10.1177/2372732215624707

Kendeou, P., & Van den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & Cognition, 35,* 1567–1577. http://dx.doi.org/10.3758/BF03193491

Kendeou, P., Walsh, E. K., Smith, E. R., & O'Brien, E. J. (2014). Knowledge revision processes in refutation texts. *Discourse Processes, 51,* 374–397. http://dx.doi.org/10.1080/0163853X.2014.913961

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31,* 338–368. http://dx.doi.org/10.3102/00028312031002338

King, A., & Rosenshine, B. (1993). Effects of guided cooperative questioning on children's knowledge construction. *Journal of Experimental Education, 61,* 127–148. http://dx.doi.org/10.1080/00220973.1993.9943857

King, A., Staffieri, A., & Adelgais, A. (1998). Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology, 90,* 134–152. http://dx.doi.org/10.1037/0022-0663.90.1.134

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95,* 163–182. http://dx.doi.org/10.1037/0033-295X.95.2.163

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, UK: Cambridge University Press.

Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71–104). Mahwah, NJ: Erlbaum.

Kintsch, W., & Mross, E. F. (1985). Context effects in word identification. *Journal of Memory and Language, 24,* 336–349. http://dx.doi.org/10.1016/0749-596X(85)90032-4

Kucan, L., & Beck, I. L. (1997). Thinking aloud and reading comprehension research: Inquiry, instruction, and social interaction. *Review of Educational Research, 67,* 271–299. http://dx.doi.org/10.3102/00346543067003271

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Long, D. L., & Chong, J. L. (2001). Comprehension skill and global coherence: A paradoxical picture of poor comprehenders' abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1424–1429. http://dx.doi.org/10.1037/0278-7393.27.6.1424

Long, D. L., Oppy, B. J., & Seely, M. R. (1994). Individual differences in the time course of inferential processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1456–1470. http://dx.doi.org/10.1037/0278-7393.20.6.1456

Markman, E. M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development, 50,* 643–655. http://dx.doi.org/10.2307/1128929

*Mccormick, S., & Hill, D. S. (1984). An analysis of the effects of two procedures for increasing disabled readers' inferencing skills. *The Journal of Educational Research, 77,* 219–226. http://dx.doi.org/10.1080/00220671.1984.10885527

McGee, A., & Johnson, H. (2003). The effect of inference training on skilled and less skilled comprehenders. *Educational Psychology, 23,* 49–59. http://dx.doi.org/10.1080/01443410303220

McKenzie, G. R. (1972). Some effects of frequent quizzes on inferential thinking. *American Educational Research Journal, 9,* 231–240. http://dx.doi.org/10.3102/00028312009002231

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38,* 1–30. http://dx.doi.org/10.1207/s15326950dp3801_1

McNamara, D. S., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14,* 1–43. http://dx.doi.org/10.1207/s1532690xci1401_1

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of Learning and Motivation, 51,* 297–384. http://dx.doi.org/10.1016/S0079-7421(09)51009-2

McNamara, D. S., & McDaniel, M. A. (2004). Suppressing irrelevant information: Knowledge activation or inhibition? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 465–482. http://dx.doi.org/10.1037/0278-7393.30.2.465

*McNamara, D. S., O'Reilly, T. P., Best, R. M., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research, 34,* 147–171. http://dx.doi.org/10.2190/1RU5-HDTJ-A5C8-JVWE

McNamara, D. S., O'Reilly, T., & de Vega, M. (2007). Comprehension skill, inference making, and the role of knowledge. In F. Schmalhofer & C. A. Perfetti (Eds.), *Higher level language processes in the brain: Inference and comprehension processes* (pp. 233–251). Mahwah, NJ: Erlbaum.

Nicholas, D. W., & Trabasso, T. (1980). Toward a taxonomy of inferences for story comprehension. In F. Wilkening, J. Becker, & T. Trabasso (Eds.), *Information integration by children* (pp. 243–265). Hillsdale, NJ: Erlbaum.

Oakhill, J. V. (1984). Inferential and memory skills in children's comprehension of stories. *British Journal of Educational Psychology, 54,* 31–39. http://dx.doi.org/10.1111/j.2044-8279.1984.tb00842.x

Oakhill, J., Hartt, J., & Samols, D. (2005). Levels of comprehension monitoring and working memory in good and poor comprehenders. *Reading and Writing, 18,* 657–686. http://dx.doi.org/10.1007/s11145-005-3355-z

Oakhill, J. V., & Patel, S. (1991). Can imagery training help children who have comprehension problems? *Journal of Research in Reading, 14,* 106–115. http://dx.doi.org/10.1111/j.1467-9817.1991.tb00012.x

O'Flahavan, J. F., Hartman, D. K., & Pearson, D. P. (1989). *Teacher questioning and feedback practices after the cognitive revolution: Replication and extension of Guszak's (1967) study* (Tech. Rep. No. No. 461). University of Illinois, Champaign, IL.

Olson, M. W. (1985). Text type and reader ability: The effects on paraphrase and text-based inference questions. *Journal of Reading Behavior, 17,* 199–214. http://dx.doi.org/10.1080/10862968509547540

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43,* 121–152. http://dx.doi.org/10.1080/01638530709336895

Paris, S. C., & Lindauer, B. K. (1976). The role of inference in children's comprehension and memory. *Cognitive Psychology, 8,* 217–227. http://dx.doi.org/10.1016/0010-0285(76)90024-4

Pearson, P. D., & Dole, J. A. (1987). Explicit comprehension instruction: A review of research and a new conceptualization of instruction. *The Elementary School Journal, 88,* 151–165. http://dx.doi.org/10.1086/461530

Pearson, P. D., & Johnson, D. (1978). *Teaching reading comprehension.* New York, NY: Holt, Rinehart, & Winston.

Perfetti, C. A., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18,* 22–37. http://dx.doi.org/10.1080/10888438.2013.827687

Perfetti, C. A., & Stafura, J. (2015). Comprehending implicit meanings in text without making inferences. In E. J. O'Brien, A. E. Cook, & R. F. Lorch (Eds.), *Inferences during reading* (pp. 160–181). Cambridge, United Kingdom: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781107279186.002

*Price-Cheves, D. A. (1973). *An experiment in developing the ability to make inferences and to ask questions which require inference in sixth-grade students* (Unpublished doctoral dissertation). University of Missouri, Columbia, MO.

Raphael, T. E. (1984). Teaching learners about sources of information for answering comprehension questions. *Journal of Reading, 27,* 303–311.

Raphael, T. E., & McKinney, J. (1983). An examination of fifth- and eighth-grade children's question-answering behavior: An instructional study in metacognition. *Journal of Reading Behavior, 15,* 67–86. http://dx.doi.org/10.1080/10862968309547490

Raphael, T. E., & Wonnacott, C. A. (1985). Heightening fourth grade students' sensitivity to sources of information for answering comprehension questions. *Reading Research Quarterly, 20,* 282–296. http://dx.doi.org/10.2307/748019

Raudenbush, S. W. (1994). Random effect models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–322). New York, NY: Russell Sage Foundation.

Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology, 80,* 16–20. http://dx.doi.org/10.1037/0022-0663.80.1.16

*Reutzel, D. R., & Hollingsworth, P. M. (1988). Highlighting key vocabulary: A generative-reciprocal procedure for teaching selected inference types. *Reading Research Quarterly, 23,* 358–378. http://dx.doi.org/10.2307/748047

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General, 126,* 211–227. http://dx.doi.org/10.1037/0096-3445.126.3.211

*Seifert, T. L. (1993). Effects of elaborative interrogation with prose passages. *Journal of Educational Psychology, 85,* 642–651. http://dx.doi.org/10.1037/0022-0663.85.4.642

Seigneuric, A., & Ehrlich, M. F. (2005). Contribution of working memory capacity to children's reading comprehension: A longitudinal investigation. *Reading and Writing, 18,* 617–656. http://dx.doi.org/10.1007/s11145-005-2038-0

Sesma, H. W., Mahone, E. M., Levine, T., Eason, S. H., & Cutting, L. E. (2009). The contribution of executive skills to reading comprehension. *Child Neuropsychology, 15,* 232–246. http://dx.doi.org/10.1080/09297040802220029

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–284). New York, NY: Russell Sage Foundation.

Shadish, W. R., Robinson, L., & Congxiao, L. (1999). *ES: A computer program for effect size calculation.* Memphis, TN: University of Memphis.

Shoop, M. (1982). Improving inferential comprehension of context by combining instructional techniques. *Reading Improvement, 19,* 266–273.

Sinatra, G. M., Beck, I. L., & McKeown, M. G. (1993). How knowledge influenced two interventions designed to improve comprehension. *Reading Psychology, 14,* 141–163. http://dx.doi.org/10.1080/0270271931 40203

Singer, M. (1988). Inferences in reading comprehension. In M. Daneman, G. E. MacKinnon, & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (pp. 177–219). New York, NY: Academic Press.

Singer, M., Andruslak, P., Reisdorf, P., & Black, N. L. (1992). Individual differences in bridging inference processes. *Memory & Cognition, 20,* 539–548. http://dx.doi.org/10.3758/BF03199586

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore, MD: Johns Hopkins University Press.

*Spires, H. A., & Donley, J. (1998). Prior knowledge activation: Inducing engagement with informational texts. *Journal of Educational Psychology, 90,* 249–260. http://dx.doi.org/10.1037/0022-0663.90.2.249

Stein, B. S., Bransford, J. D., Franks, J. J., Owings, R. A., Vye, N. J., & McGraw, W. (1982). Differences in the precision of self-generated elaborations. *Journal of Experimental Psychology: General, 111,* 399–405. http://dx.doi.org/10.1037/0096-3445.111.4.399

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54,* 1046–1055. http://dx.doi.org/10.1016/S0895-4356(01)00377-8

*Stitt, J. H. (1967). *Effects of instruction on children's inferential thinking* (Unpublished doctoral dissertation). University of California, CA, Los Angeles.

*Sundbye, N. (1987). Text explicitness and inferential questioning: Effect on story understanding and recall. *Reading Research Quarterly, 22,* 82–98. http://dx.doi.org/10.2307/747722

Sutton, A. J. (2009). Publication bias. In H. Cooper, H. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 435–452). New York, NY: Russell Sage Foundation.

Swanson, H. L., & Berninger, V. (1995). The role of working memory in skilled and less skilled readers' comprehension. *Intelligence, 21,* 83–108. http://dx.doi.org/10.1016/0160-2896(95)90040-3

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language, 24,* 612–630. http://dx.doi.org/10.1016/0749-596X(85)90049-X

van den Broek, P., Kendeou, P., Kremer, K., Lynch, J., Butler, J., & White, M. J. (2005). Assessment of comprehension abilities in young children. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 107–129). Mahwah, NJ: Erlbaum.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York, NY: Academic Press.

Weekes, B. S., Hamilton, S., Oakhill, J. V., & Holliday, R. E. (2008). False recollection in children with reading comprehension difficulties. *Cognition, 106,* 222–233. http://dx.doi.org/10.1016/j.cognition.2007.01.005

*Wells, S. (1986). *Effects of guided instruction to improve use of inferences by the junior high reader* (Unpublished doctoral dissertation). University of Northern Colorado, Fort Collins, CO.

What Works Clearinghouse. (2014). *What Works Clearinghouse Procedures and Standards Handbook version 3.0.* Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

Williams, J. P. (1993). Comprehension of students with and without learning disabilities: Identification of narrative themes and idiosyncratic text representations. *Journal of Educational Psychology, 85,* 631–641. http://dx.doi.org/10.1037/0022-0663.85.4.631

Wilson, M. M. (1979). The processing strategies of average and below average readers answering factual and inferential questions of three equivalent passages. *Journal of Reading Behavior, 11,* 235–245. http://dx.doi.org/10.1080/10862967909547327

*Winne, P. H., Graham, L., & Prock, L. (1993). A model of poor readers' text-based inferencing: Effects of explanatory feedback. *Reading Research Quarterly, 28,* 52–66. http://dx.doi.org/10.2307/747816

Wolfe, M. B. (2005). Memory for narrative and expository text: Independent influences of semantic associations and text organization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 359–364. http://dx.doi.org/10.1037/0278-7393.31.2.359

*Yuill, N., & Joscelyne, T. (1988). Effect of organizational cues and strategies on good and poor comprehenders' story understanding. *Journal of Educational Psychology, 80,* 152–158. http://dx.doi.org/10.1037/0022-0663.80.2.152

*Yuill, N., & Oakhill, J. (1988). Effects of inference awareness training on poor reading comprehension. *Applied Cognitive Psychology, 2,* 313–345. http://dx.doi.org/10.1002/acp.2350020105

Yuill, N., & Oakhill, J. (1988). Effects of inference awareness training on poor reading comprehension. *Applied Cognitive Psychology, 2,* 33–45. http://dx.doi.org/10.1002/acp.2350020105

Yuill, N., & Oakhill, J. (1991). *Children's problems in text comprehension: An experimental investigation.* Cambridge, United Kingdom: Cambridge University Press.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 386–397. http://dx.doi.org/10.1037/0278-7393.21.2.386

Appendix A

Study Characteristics

Table A1

General and Methodological Characteristics of Studies

	k (%)
Publication type	
Journal	27 (82%)
Dissertation	4 (12%)
Book	2 (6%)
Publication year	
1960–1969	1 (3%)
1970–1979	1 (3%)
1980–1989	21 (64%)
1990–1999	7 (21%)
2000–2009	2 (6%)
2010–2013	1 (3%)
Type of control	
Business as usual control	8 (24%)
BAU with material exposure or lesser treatment	25 (76%)
Assignment	
Random/matched	26 (79%)
Quasi-experimental	7 (21%)
Attrition	
Less than 15%	32 (97%)
More than 16%	1 (3%)

Note. BAU = Business as usual control.

Table A2

Measurement Characteristics of Studies

Measurement format	Open-ended or recall k (%)	Multiple choice k (%)	Mixed format k (%)
General (k = 13)	8 (62%)	5 (38%)	
Inference (k = 25)	20 (80%)	4 (16%)	2 (8%)
Literal (k = 18)	15 (83%)	3 (17%)	

	Standardized k (%)	Custom k (%)
Measure type		
General (k = 13)	7 (54%)	6 (46%)
Inference (k = 25)	2 (8%)	23 (92%)
Literal (k = 18)	1 (6%)	17 (94%)

	Listening k (%)	Reading k (%)
Measure presentation		
General (k = 13)	3 (23%)	10 (77%)
Inference (k = 25)	6 (24%)	19 (76%)
Literal (k = 18)	3 (17%)	15 (83%)

	Inter-rater k (%)	Internal consistency k (%)	Not reported k (%)
Measure reliability			
Across studies (k = 56)	30 (55%)	14 (24%)	12 (21%)

(*Appendices continue*)

Table A3
Contextual Intervention Characteristics of Studies

	k (%)
Group format	
One to one	9 (27%)
Small group	6 (18%)
Whole class	18 (55%)
Instruction intensity in hours	
1–5	16 (48%)
6–10	8 (24%)
11–15	2 (6%)
16–20	4 (12%)
21+	3 (9%)
Inference type	
Text-based	9 (27%)
Elaborative or both	24 (73%)
Instructor	
Teacher	14 (42%)
Researcher	16 (48%)
Self-directed	2 (6%)
Parent	1 (3%)
Text type	
Narrative	15 (45%)
Expository	10 (30%)
Both	5 (15%)
Not reported	3 (9%)
Implementation fidelity	
Yes	11 (33%)
Not reported	22 (66%)
Instructor training	
Yes	12 (36%)
Not reported	21 (64%)

Table A4
Participant Characteristics

	k (%)
Grade	
K–2[a]	6 (18%)
3–5	13 (39%)
6–8	12 (36%)
9–12[a]	2 (6%)
Reading ability	
Skilled	23 (69%)
Less skilled	10 (31%)
Socioeconomic status	
Low	5 (15%)
Middle/high	18 (55%)
Not reported	10 (30%)

[a] No studies were conducted in kindergarten or Grades 1, 10, 11, and 12.

(*Appendices continue*)

Appendix B

Additional Information for Consideration of Publication Bias

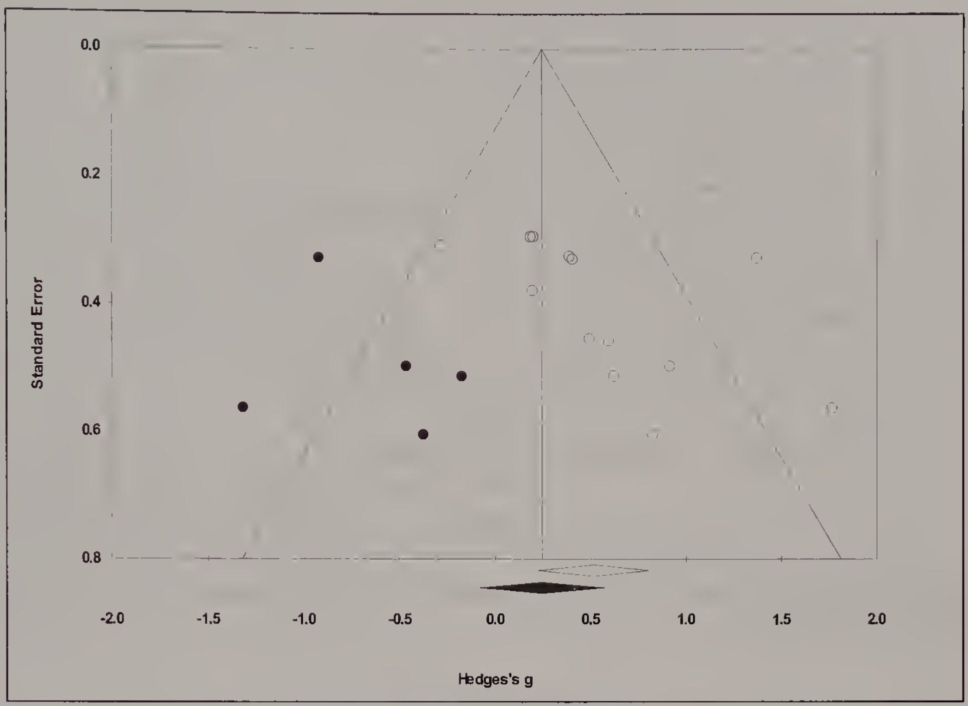


Figure B1. Funnel plot of observed (k = 13) and imputed (k = 5) general comprehension effects by *SE* using a 95% confidence interval (CI).

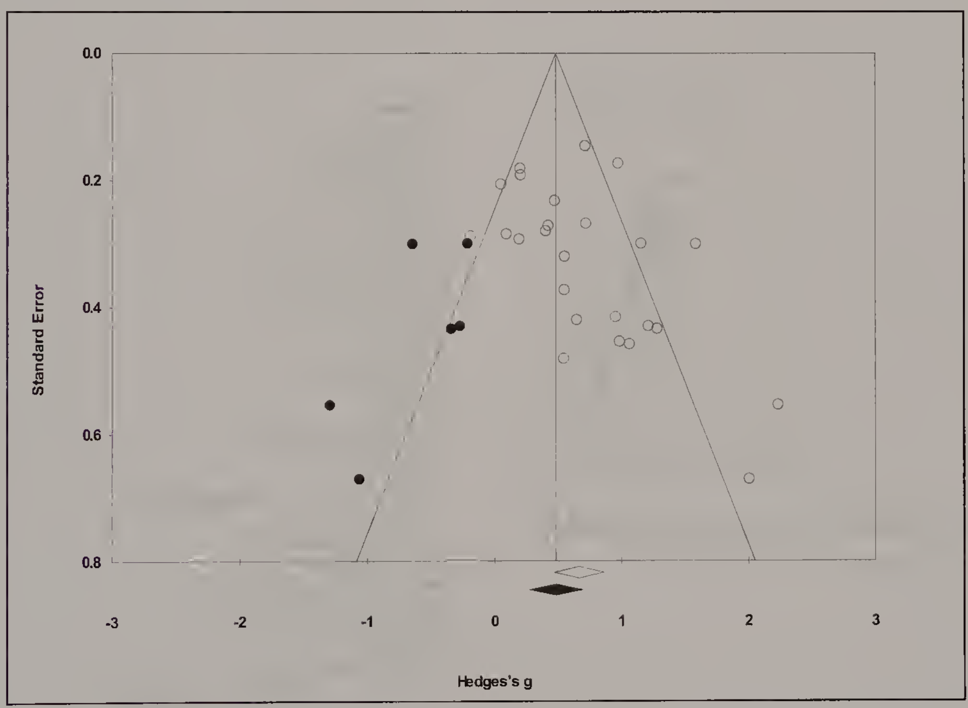


Figure B2. Funnel plot of observed (k = 25) and imputed (k = 6) inferential comprehension effects by *SE* using a 95% confidence interval (CI).

(*Appendices continue*)

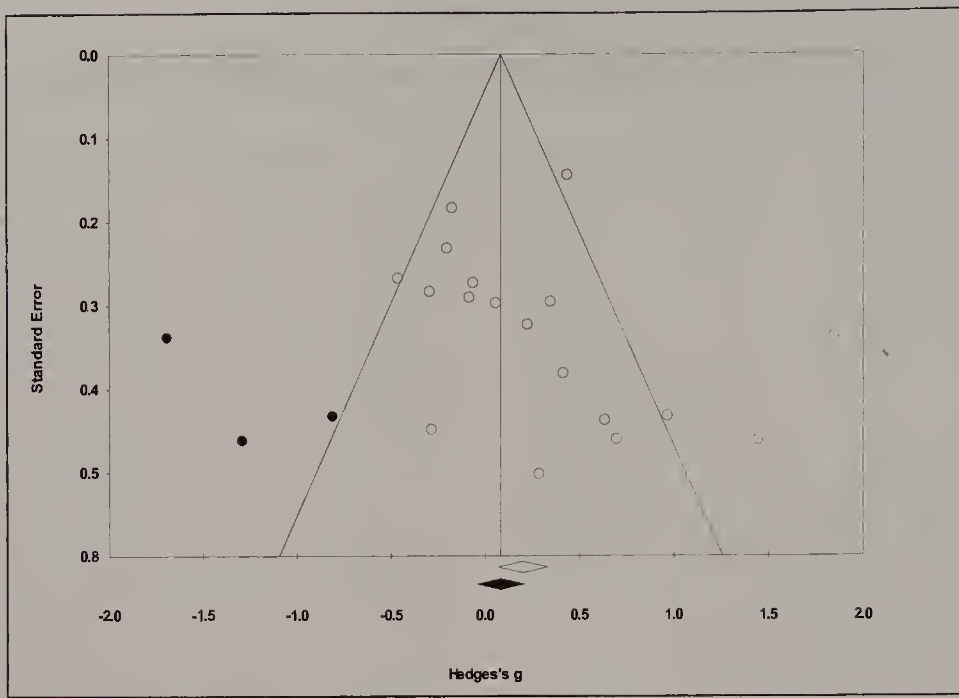


Figure B3. Funnel plot of observed (k = 18) and imputed (k = 3) literal comprehension effects by *SE* using a 95% confidence interval (CI).

Appendix C

Funnel Plots

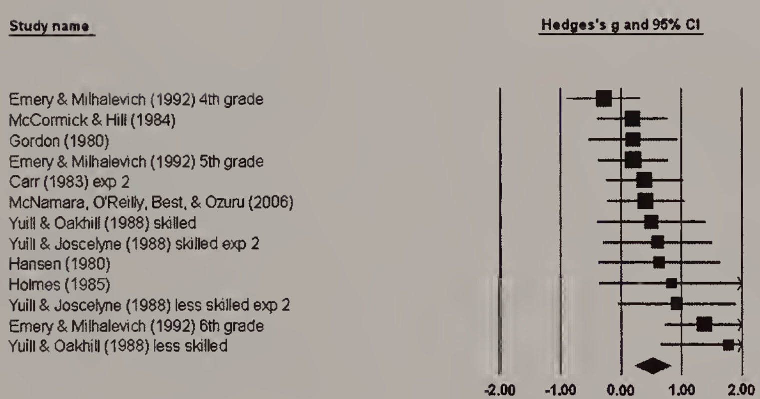


Figure C1. Study characteristics. Forest plot of general comprehension effects using a 95% confidence interval.

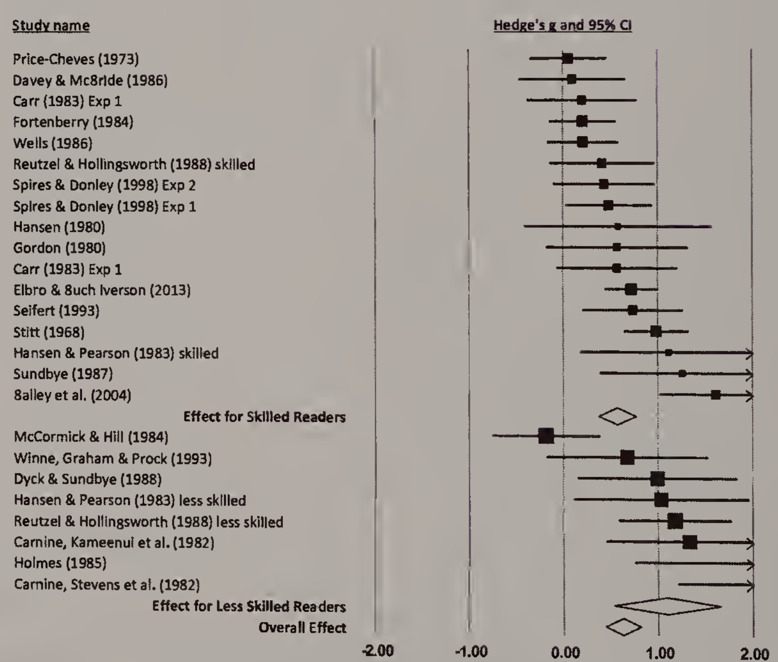


Figure C2. Forest plots of main effects. Forest plot of a mixed-weight inference effects by reading ability using a 95% confidence interval.

(*Appendices continue*)

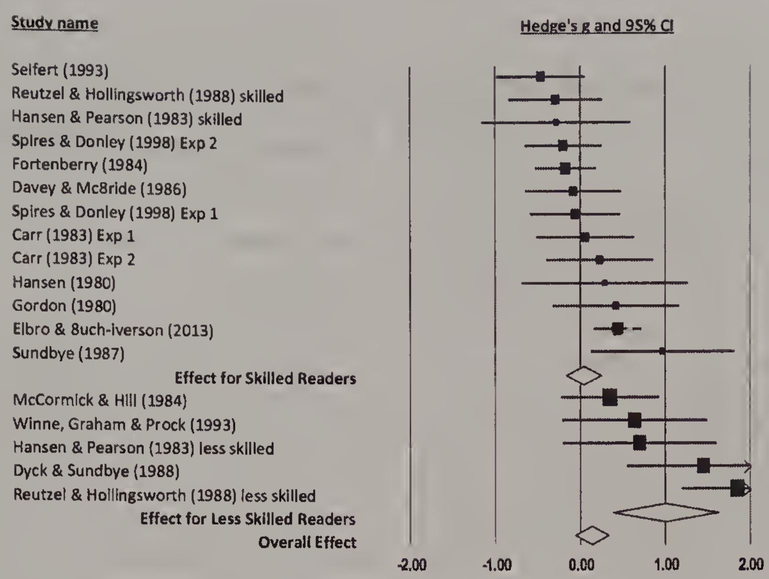


Figure C3. Additional information for consideration of publication bias. Forest plot of mixed-weight literal effects by reading ability using a 95% confidence interval.

Received January 13, 2016
Revision received November 18, 2016
Accepted November 21, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 6, 782–793

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000179

Language-Independent and Language-Specific Aspects of Early Literacy: An Evaluation of the Common Underlying Proficiency Model

J. Marc Goodrich and Christopher J. Lonigan
Florida State University

According to the common underlying proficiency model (Cummins, 1981), as children acquire academic knowledge and skills in their first language, they also acquire language-independent information about those skills that can be applied when learning a second language. The purpose of this study was to evaluate the relevance of the common underlying proficiency model for the early literacy skills of Spanish-speaking language-minority children using confirmatory factor analysis. A total of 858 Spanish-speaking language-minority preschoolers (mean age = 60.83 months; 50.2% female) participated in this study. Results indicated that bifactor models that consisted of language-independent as well as language-specific early literacy factors provided the best fits to the data for children's phonological awareness and print knowledge skills. Correlated factors models that included skills specific to only Spanish and English provided the best fits to the data for children's oral language skills. Children's language-independent early literacy skills were significantly related across constructs and to language-specific aspects of early literacy. Language-specific aspects of early literacy skills were significantly related within but not across languages. These findings suggest that language-minority preschoolers have a common underlying proficiency for code-related skills but not language-related skills that may allow them to transfer knowledge across languages.

Keywords: language-minority, early literacy, phonological awareness, print knowledge, oral language

Early literacy skills are the developmental precursors to conventional reading skills and are measurable during the preschool years, prior to the beginning of formal reading instruction. Research has indicated that three early literacy skills are the strongest predictors of children's future reading ability: phonological awareness, print knowledge, and oral language (Lonigan, Schatschneider, & Westberg, 2008; Whitehurst & Lonigan, 1998). Phonological awareness refers to the ability to detect and manipulate the individual sound components of words, independent of meaning. Print knowledge refers to children's knowledge of the conventions of print (e.g., text is read from left to right in English) as well as knowledge of letters and letter–sound correspondence. Oral language refers to the ability to use spoken language to understand and convey meaning, and it includes children's vocabulary and syntactic knowledge, among other skills. Phonological awareness and print knowledge are code-related skills that are highly related to children's later decoding (i.e., word reading) abilities, whereas oral language is more strongly related to children's later reading comprehension skills (e.g., Storch & Whitehurst, 2002). Evidence has indicated that early literacy skills are causally related to children's later reading abilities (e.g., Byrne & Fielding-Barnsley, 1995; Hulme, Bowyer-Crane, Carroll, Duff, & Snowling, 2012). Consequently, it may be important to identify children with poor early literacy skills and intervene early to prevent difficulties in acquiring conventional reading abilities.

Children whose home language is different from that spoken by the majority of the population of the country in which they live are often referred to as language-minority (LM) children (e.g., August, Shanahan, & Escamilla, 2009). Children who speak Spanish at home comprise the largest group of LM children in the United States, and these children are at a high risk for struggling academically (Hemphill, Vanneman, & Rahman, 2011). Prior research on the early literacy skills of LM children has indicated that the same skills that are important precursors to conventional reading skills among monolingual children are also predictive of LM children's later reading abilities (e.g., Manis, Lindsey, & Bailey, 2004). However, LM children often score lower on measures of early literacy and enter elementary school with weaker English reading abilities than do monolingual children (Hoff, 2013; Lonigan, Farver, Nakamoto, & Eppe, 2013). Additionally, evidence has indicated that rates of growth of reading abilities do not differ for LM and monolingual children (Kieffer & Vukovic, 2013), indicating that the gap between LM and monolingual children does not begin to narrow once LM children are exposed to formal reading instruction in English. Therefore, it is important to understand how

This article was published Online First February 6, 2017.

J. Marc Goodrich and Christopher J. Lonigan, Department of Psychology and the Florida Center for Reading Research, Florida State University.

This research and report was supported by grants from the U.S. Department of Education's Institute of Education Sciences (R305F100027) and the Eunice Kennedy Schriver National Institute of Child Health and Human Development (HD060292). The views expressed herein are those of the authors and have not been reviewed or approved by the granting agencies.

Correspondence concerning this article should be addressed to J. Marc Goodrich, who is now at the Department of Special Education & Communication Disorders, University of Nebraska, 301 Barkley Memorial Center, Lincoln, NE 68583-0738 or to Christopher J. Lonigan, Department of Psychology, Florida State University, 1107 West Call Street, Tallahassee FL 32306-4301. E-mail: marc.goodrich@unl.edu or lonigan@psy.fsu.edu

LM children's early literacy skills develop to prevent them from falling behind their monolingual peers early in life.

The Common Underlying Proficiency Model

Theory regarding the development of language and literacy skills among LM children suggests that children can transfer knowledge across languages. According to the developmental interdependence hypothesis (Cummins, 1979), for children exposed to a second language (L2), development of that language is dependent on the level of proficiency in their first language (L1) at the time of sustained exposure to L2. This claim suggests that children with low levels of proficiency in Spanish may have more difficulties with English acquisition than would children with high levels of proficiency in Spanish. Similarly, some researchers have suggested that children with higher levels of proficiency in L1 are more likely to benefit from L2 instruction than are children with lower levels of proficiency in L1, because they can transfer preexisting L1 knowledge to their L2 when exposure to L2 begins (e.g., Cummins, 2008). However, current evidence for cross-language transfer has been predominantly correlational (e.g., Durgunoğlu, Nagy, & Hancin-Bhatt, 1993; Lindsey, Manis, & Bailey, 2003), and significant cross-language correlations of L1 and L2 skills do not necessarily represent evidence of transfer, because such findings are open to alternative explanations, such as common language-learning environments across L1 and L2. Additional research is needed to determine empirically whether cross-language transfer occurs and how to best leverage LM children's existing L1 skills when they begin to learn L2.

Cummins (1981) introduced the idea of a common underlying proficiency to describe a potential mechanism through which cross-language transfer could occur. According to the common underlying proficiency model, proficiencies in L1 and L2 are not separate abilities. Although there are surface features of each language that are distinct, L1 and L2 are intrinsically connected. As proficiency in one language develops, so does language-independent knowledge (i.e., the common underlying proficiency) that supports the development of skills in both languages. Exposure to either L1 or L2 will contribute to the development of the common underlying proficiency, which may allow children to transfer knowledge across languages. Although this model was developed based on evidence of the transferability of language-independent skills (e.g., inferring meaning from text), Cummins argued that even when a task seems relatively language-specific (e.g., spelling), there will be strong relations between L1 and L2 due to the common underlying proficiency. Some support for the common underlying proficiency model has come from studies demonstrating that educational curricula that incorporate instruction in both languages have larger effects on academic outcomes than do curricula that provide instruction exclusively in L2 (e.g., Cheung & Slavin, 2012). For many LM children in the United States, substantial exposure to L2 (i.e., English) begins with enrollment in preschool. A few studies evaluating the effectiveness of interventions for LM preschoolers' literacy-related skills have provided support for the common underlying proficiency model (e.g., Farver, Lonigan, & Eppe, 2009). However, no study to date has attempted to evaluate empirically whether LM children's L1 and L2 literacy-related skills are represented by a common underlying proficiency.

Cross-Language Relations of Early Literacy Skills

The common underlying proficiency model suggests that children's early literacy skills are related across languages. Over the past several decades, Cummins's (1981) theory has led to a large amount of research examining the cross-language relations of various academic skills among LM children (e.g., Melby-Lervåg & Lervåg, 2011). However, the common underlying proficiency may be more relevant for some skills than it is for others. Some skills—such as phonological awareness, which requires manipulation of sounds of words, independent of meaning—may be relatively language-independent. Once the understanding that words are made up of sounds that can be manipulated (e.g., isolated, removed) is acquired, children should be able to apply this skill to words they do not know and potentially to words in another language. In contrast, other skills, such as vocabulary knowledge, are more specific to a particular language. For language-independent skills, LM children should develop a common underlying proficiency that can then be applied to other languages. For language-specific skills, children should acquire knowledge that is not necessarily applicable to other languages, limiting the extent to which acquisition of knowledge and skills would be associated with development of a common underlying proficiency for those skills.

Phonological awareness is a relatively language-independent ability. For example, the knowledge that words are made up of smaller units of sound that can be manipulated is applicable to both English and Spanish, as long as the sounds of both languages can be detected. Therefore, the common underlying proficiency may be particularly relevant for the development of phonological awareness. Research has demonstrated that LM children's phonological awareness skills are significantly related across languages (e.g., Branum-Martin et al., 2006; Durgunoğlu et al., 1993). Melby-Lervåg and Lervåg (2011) conducted a meta-analysis of the correlational studies of cross-language relations of literacy-related skills and reported that the average cross-language correlation of phonological awareness was large ($r = .54$). If this cross-language correlation is due to a common underlying proficiency for phonological awareness, it would be expected that development of phonological awareness in L1 would support the development of phonological awareness in L2, and vice versa.

Print knowledge is somewhat less language-independent than is phonological awareness. Although print knowledge consists of the language-independent knowledge that letters have names and are associated with sounds, it also includes language-specific information (e.g., specific letter names and letter–sound correspondences). Therefore, the common underlying proficiency model may be less relevant for the development of print knowledge than it is for phonological awareness. The extent to which LM children develop a common underlying proficiency for print knowledge and transfer print knowledge skills across languages should be dependent on the amount of overlap in information pertaining to print knowledge across two languages. For example, because Spanish and English share almost all alphabetic symbols and letter names are similar across the two languages, the common underlying proficiency model may be more relevant for Spanish and English print knowledge than it is for two languages that have fewer similarities in surface-level features (e.g., English and Arabic). Bialystok, Luk, and Kwan (2005) reported that, consistent with

this idea, L1 and L2 were related only when children were acquiring two languages that used the same writing system. Several studies have reported significant cross-language correlations of print knowledge among Spanish-speaking LM children (e.g., Goodrich, Lonigan, & Farver, 2013; Lindsey et al., 2003), indicating that the common underlying proficiency model may be relevant for print knowledge.

Oral language consists of primarily language-specific knowledge. For example, oral language skills of preschool children are commonly assessed using vocabulary measures. With the exception of cognates, vocabulary knowledge is language-specific because words in a language are arbitrarily associated with their underlying concepts. Therefore, the common underlying proficiency model should have limited relevance for the development of oral language among preschool children. For example, although children may have language-independent knowledge of a concept because they know the corresponding word for that concept in L1, there is often little to no information about that concept or its L1 label that children could use to acquire the word in L2. Results of prior research have indicated that LM children's vocabularies are distributed across their two languages, with approximately 70% of words known in L1 or L2 but not both (Peña, Bedore, & Zlatic-Giunta, 2002). Additionally, several studies have indicated that cross-language correlations of vocabulary knowledge are often nonsignificant or negative (Bialystok et al., 2005; Goodrich, Lonigan, Kleuver, & Farver, 2016; Gottardo & Mueller, 2009), suggesting that there is not a common underlying proficiency for young LM children's oral language skills; however, in their meta-analysis Melby-Lervåg and Lervåg (2011) reported that the correlation between L1 and L2 oral language skills was significant, albeit small (r = .16), suggesting that a common underlying proficiency may play a small role in the development of L1 and L2 oral language.

The Current Study

The purpose of this study was to evaluate the relevance of the common underlying proficiency model for the development of early literacy skills among Spanish-speaking LM preschoolers by using bifactor models. Bifactor models are a special case of confirmatory factor analysis in which variance in indicators is partitioned into general variance that is common across all indicators (i.e., all items load onto a general factor) as well as construct-specific variance (i.e., items also load onto construct-specific factors; Reise, 2012). Bifactor models account for overlapping variance across constructs (i.e., the general factor). Therefore, if two constructs are not significantly related to each other, a bifactor model should not provide an improvement in fit to the data over a more parsimonious correlated-factors model. In this study, we estimated bifactor models to determine the extent to which variance in items on Spanish and English early literacy assessments is shared across languages or is unique to the language of assessment. Evidence that a bifactor model fit the data significantly better than did a correlated-factors model would indicate that children have a common underlying proficiency for early literacy skills (as represented by the general factor). It was hypothesized that, based on theory and prior evidence (e.g., Melby-Lervåg & Lervåg, 2011), the common underlying proficiency would be relevant for phonological awareness and print knowledge but not for oral language. Additionally, it was expected that the common underlying proficiency model would be more relevant for phonological awareness than it would for print knowledge.

We also evaluated cross-construct correlations between language-specific early literacy abilities and LM children's common underlying proficiencies for early literacy skills. We expected that unique Spanish abilities would not be significantly related to unique English abilities. If children's common underlying proficiencies for early literacy abilities represented variance unique to each construct, it would not be expected that the common underlying proficiencies would be related to language-specific aspects of early literacy across constructs. For example, if language-independent phonological awareness and print knowledge skills represented abilities that were entirely unique to phonological awareness and print knowledge, respectively, those constructs would not be related to each other or to other language-specific aspects of early literacy (e.g., English oral language). However, some researchers have speculated that evidence of cross-language relations of academic skills emerges due to children's underlying language-learning capacity (Castilla, Restrepo, & Perez-Leroux, 2009). If the common underlying proficiency is indicative of a general language-learning capacity and is not unique to any specific construct, it would be expected that children's common underlying proficiencies would be related to each other and to language-specific aspects of early literacy across constructs (e.g., common underlying proficiency for print knowledge would be correlated with Spanish-specific aspects of phonological awareness).

Method

Participants

Spanish-speaking LM children (N = 858) enrolled in 102 preschool classrooms participated in this study. Children in this study represented the LM portion of a larger sample recruited for participation in a curriculum evaluation study that was designed to target the development of early literacy skills in at-risk, low-income preschool children. Children came from classrooms in which at least 50% of children were Spanish-speaking LM children, and all preschool classrooms were required to have at least one teacher who was a fluent Spanish speaker. Children in this sample were recruited from several regions across the United States, including Central Florida, South Florida, New Mexico, Kansas, and Southern California. Consequently, children's home language experiences came from diverse countries of origin, including Mexico, Cuba, Puerto Rico, and Central and South American countries. Children ranged in age from 44 months to 74 months (M = 60.83, SD = 4.74). Among participants for whom data on sex were available (n = 804), 49.8% of participants were identified as male. Among participants for whom parent report data were available, parent report of language spoken at home indicated that for 71.7% of children Spanish was the language most frequently spoken at home, for 13.9% of children English was the language most frequently spoken at home, for 12.6% of children Spanish and English were spoken equally often at home, and for 1.8% of children some other pattern of languages was spoken at home. Among participants for whom parent report data were available, 22.4% of mothers and 19.1% of fathers were born

in the United States or Puerto Rico, whereas 94.5% of children were born in the United States or Puerto Rico, indicating that the majority of these children were first generation.

Measures

Phonological awareness. Children completed the Phonological Awareness subtest of the Test of Preschool Early Literacy (TOPEL; Lonigan, Wagner, Torgesen, & Rashotte, 2007). This subtest contains 15 blending and 12 elision items. Blending items require children to combine words or parts of words to form a new word (e.g., combining *star* and *fish* to form *starfish*). Elision items require children to remove individual sounds or segments of sound from words to form new words (e.g., removing *flower* from *sunflower* to create *sun*). Six of the blending items are multiple choice, and nine are free response. Six of the elision items are multiple choice, and six are free response. Multiple-choice items require children to either point to a visual depiction of the correct answer (out of four possible pictures) or verbally say the correct answer. Free-response items require children to verbally say the correct answer in the absence of pictures. Items on the Phonological Awareness subtest span the range of linguistic complexity, with items requiring manipulation of individual phonemes, syllables, and whole words. Internal consistency reliability on the Phonological Awareness subtest of the TOPEL is high ($\alpha = .89$). Children also completed the Blending and Elision subtests of the Spanish Preschool Early Literacy Assessment (SPELA; Lonigan, 2012). The Blending and Elision subtests of the SPELA contain 32 items each, 16 of which are multiple choice and 16 of which are free response. The SPELA is designed to mirror the TOPEL in structure and form. Internal consistency reliability on the SPELA was very high in this sample of children (for Blending $\alpha = .96$; for Elision $\alpha = .93$).

Print knowledge. Children completed the Print Knowledge subtests of the TOPEL and the Spanish Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP-Spanish; Lonigan, Farver, & Eppe, 2002). The Print Knowledge subtests each contain 36 items. On the TOPEL, the Print Knowledge subtest contains four items each for print concepts (e.g., "Which picture shows the name of the book?"), letter discrimination (e.g., "Which is a letter?"), and word discrimination (e.g., "Which can you read?"), all of which are multiple choice. Sixteen items assess children's knowledge of letter names (e.g., "Which one is 'b'?"), six of which are multiple choice and 10 of which are free response. The remaining eight items assess children's knowledge of letter–sound correspondence (e.g., "Which one makes the /b/ sound?"), four of which are multiple choice and four of which are free response. Internal consistency reliability for the Print Knowledge subtest of the TOPEL is very high ($\alpha = .96$). The Print Knowledge subtest of the Pre-CTOPPP-Spanish is a direct Spanish-language translation of the TOPEL Print Knowledge subtest. Internal consistency reliability for the Print Knowledge subtest of the Pre-CTOPPP Spanish was very high in this sample ($\alpha = .93$).

Oral language. Children completed the Definitional Vocabulary subtests of the TOPEL and Pre-CTOPPP-Spanish. The Definitional Vocabulary subtests contain 35 items, each of which have an expressive and a definitional component. For the expressive component of the item, children are asked to name a picture (e.g., "What is this?"). For the definitional component of the item, children are asked a follow-up question that requires them to describe a feature or function of the item (e.g., "What is it for?"). Internal consistency reliability for the Definitional Vocabulary subtest of the TOPEL is very high ($\alpha = .95$). The Definitional Vocabulary subtest of the Pre-CTOPPP-Spanish is a direct Spanish-language translation of the TOPEL Definitional Vocabulary subtest. Internal consistency reliability for the Definitional Vocabulary subtest of the Pre-CTOPPP-Spanish was very high in this sample ($\alpha = .98$).

Procedure

Written informed consent was obtained from parents or guardians of participants prior to data collection. All assessments were completed in a quiet area of the children's preschool by trained bilingual research assistants. Order of completion of Spanish and English measures varied across participants, and Spanish and English assessments were completed on separate days that were no more than 1 week apart. Answers were coded as correct only if they were given in the language being assessed.

Data Analysis

Confirmatory factor analysis was conducted using Mplus Version 7.31 (Muthén & Muthén, 1998–2015) using full information maximum likelihood estimation to account for missing data. For Spanish variables, no variable had more than 5.1% missing data (print knowledge). For English definitional vocabulary, there was a larger amount of missing data (14.9%). However, there was less missing data for English phonological awareness (7.8%) and print knowledge (.6%). For each outcome (i.e., blending, elision, print knowledge, expressive vocabulary, definitional vocabulary), categorical, item-level data were analyzed to determine the factor structure of children's Spanish and English early literacy skills. A one-factor model in which all items for a construct (i.e., phonological awareness, print knowledge, oral language) loaded onto the same factor was estimated. The one-factor model was then compared to a two-factor model in which Spanish items loaded onto a Spanish factor and English items loaded onto an English factor. The two-factor model was then compared to a bifactor model in which Spanish items loaded onto a specific Spanish factor, English items loaded onto a specific English factor, and all items also loaded onto a general factor for the skill being assessed. In the bifactor model, an orthogonality constraint was imposed on all factors such that correlations between all factors were fixed to zero. Model comparisons were done using the likelihood ratio test, as well as comparing the values of Akaike's information criterion (AIC) and the sample-size adjusted Bayesian information criterion (ABIC). A significant likelihood ratio test indicates better fit for the less parsimonious model, and a decrease greater than 10 in AIC or ABIC values represents significant improvement in model fit (Kass & Raftery, 1995). For some model comparisons, the likelihood ratio test resulted in a negative test statistic. In these instances, the strictly positive likelihood ratio test was used (Asparouhov & Muthén, 2013). Confirmatory factor analysis was conducted using the maximum likelihood estimator with robust standard

errors. All factor loadings were freely estimated, and the variances of the factors were fixed for scale dependency. To determine variance accounted for by each factor, we computed omega, omega hierarchical, and omega subscale. These statistics can be used as estimates of reliability of factors in bifactor models and as metrics of variance accounted for by each factor (details of the computation of various forms of omega are described in Reise, Bonifay, & Haviland, 2013).

All models included a sandwich estimator to adjust the standard errors to account for the nested structure of the data (i.e., children nested within classrooms). Although classrooms were nested within state, state-level variance components computed for all summary variables were not statistically significant (all *ps* $> .20$), whereas all classroom-level variance components were statistically significant (all *ps* $< .001$). For Spanish-language variables, classroom-level intraclass correlations (ICCs) ranged from .28 to .36. For English-language variables, classroom-level ICCs ranged from .19 to .29.

Results

Descriptive statistics for raw scores on the Spanish and English early literacy measures are reported in Table 1. Standard scores were computed for English early literacy skills; however, standard scores are not available for the measures of Spanish early literacy. Standard scores for English phonological awareness were based on the combined scores for all blending and elision items, and standard scores for English definitional vocabulary were based on the combined scores for all expressive and definitional vocabulary items. Children's English phonological awareness and print knowledge skills were in the average range (for phonological awareness, $M = 94.66$, $SD = 17.52$; for print knowledge, $M = 103.20$, $SD = 14.12$), despite having below-average English language skills ($M = 85.39$, $SD = 18.32$).

Phonological Awareness

Although blending and elision items measure the same underlying phonological awareness construct, within each language a two-factor model of phonological awareness in which items loaded onto separate Blending and Elision factors provided significantly better fit to the data than did a one-factor model in which all items loaded onto the same Phonological Awareness factor (likelihood ratio test for Spanish = 319.04, $p < .001$; likelihood ratio test for English = 575.16, $p < .001$). The Blending and Elision factors were significantly correlated for both Spanish and English phonological awareness (for Spanish: $r = .72, p < .001$; for English: $r = .64$, $p < .001$). Because Blending and Elision were separable factors, subsequent models evaluated the measurement structure of English and Spanish phonological awareness separately for blending and elision items.

A summary of model fit statistics is reported in Table 2. For blending, the two-factor model provided significantly better fit to the data than did the one-factor model, according to the likelihood ratio test and other model fit indices (i.e., AIC, ABIC). The correlation between the English and Spanish Blending factors was statistically significant ($r = .31$, $p < .001$). The bifactor model provided significantly better fit to the data than did the two-factor model. Detailed results of the bifactor model are reported in Table 3. In this model, all English and Spanish blending items measured variance in blending ability that was unique to English and Spanish, respectively. All Spanish items and the majority of English items also measured variance in blending ability that was shared across languages. Omega statistics for the blending model are reported in the leftmost column of Table 4. Omega statistics indicated that approximately 35% of variance in total blending scores was due to the General factor, approximately 53% was due to the Spanish factor, and approximately 12% was due to the English factor. Additionally, 28% of variance in scores on Spanish blending items was due to the General factor, and 72% was due to the Spanish factor. Similarly, 14% of variance in scores on English blending items was due to the General factor, and 86% was due to the English factor.

For elision, the two-factor model provided significantly better fit to the data than did the one-factor model. The correlation between the Spanish and English Elision factors was statistically significant ($r = .37$, $p < .001$). The bifactor model provided significantly better fit to the data than did the two-factor model. Detailed results of the bifactor model are reported in Table 5. All English items measured variance in elision ability that was unique to English, and English free-response items measured variance that was common across languages. All Spanish items measured variance in elision ability that was unique to Spanish, and Spanish free-response items also measured variance that was common across languages; however, the majority of multiple choice items measured only variance in elision ability that was specific to Spanish or English.

The formulas for the computation of omega statistics are valid only when all factor loadings are positive (Gignac, 2014). Therefore, omega for the bifactor elision model was estimated with negative loadings from the full model removed to compute estimates of variance accounted for by the bifactor model. Omega statistics for the elision model are reported in the middle column of Table 4. Omega statistics indicated that approximately 45% of variance in total elision scores was due to the General factor, approximately 46% was due to the Spanish factor, and approximately 9% was due to the English factor. Additionally, 44% of variance in scores on Spanish elision items was due to the General factor, and 56% was due to the Spanish factor. In contrast, 5% of variance in scores on English elision items was due to the General factor, and 95% was due to the English factor.

Table 1
Descriptive Statistics of Preschoolers' English and Spanish Early Literacy Skills

	English			Spanish		
Variable	Range	*M*	*SD*	Range	*M*	*SD*
Blending	0–15	11.45	3.74	0–32	21.67	9.77
Elision	0–12	7.30	3.24	0–32	14.94	8.65
Print	0–36	24.75	10.12	0–36	20.29	8.74
Expressive vocabulary	0–35	23.05	7.12	0–35	14.31	9.68
Definitional vocabulary	0–35	22.55	8.05	0–35	14.87	10.55

Note. Data are raw scores.

Table 2
Model Fit Statistics From Confirmatory Factor Analysis of Spanish and English Phonological Awareness, Print Knowledge, and Vocabulary

Construct model	Log likelihood	AIC	ABIC	Parameters	Difference test
Blending					
One factor	−16,319.19	32,826.38	32,971.79	94	
Two factor	−15,199.01	30,588.02	30,734.99	95	3,230.50***
Bifactor	−14,684.99	29,651.99	29,870.11	141	772.32***
Elision					
One factor	−18,172.21	36,520.43	36,656.56	88	
Two factor	−17,487.84	35,153.69	35,291.37	89	9,046.54***
Bifactor	−16,920.71	34,105.43	34,309.63	132	1,226.36***
Print knowledge					
One factor	−27,522.89	55,333.78	55,556.54	144	
Two factor	−26,246.81	52,783.61	53,007.92	145	340.53***
Bifactor	−24,390.91	49,213.82	49,547.96	216	4,578.71***
Expressive vocabulary					
One factor	−23,763.92	47,807.85	48,024.76	140	
Two factor	−21,404.56	43,091.12	43,309.59	141	647.58***
Definitional vocabulary					
One factor	−27,267.66	54,815.33	55,032.24	140	
Two factor	−24,779.49	49,840.97	50,059.43	141	44,994.20***

Note. AIC = Akaike's information criterion; ABIC = sample-size-adjusted Bayesian information criterion.
*** $p < .001$.

Print Knowledge

A summary of model fit statistics for Spanish and English print knowledge is reported in Table 2. The two-factor model provided significantly better fit to the data than did the one-factor model. The correlation between the Spanish and English Print Knowledge factors was high ($r = .61$, $p < .001$). The bifactor model provided significantly better fit to the data than did the two-factor model. Factor loadings from the bifactor model are reported in Table 6. All English print knowledge items and the majority of Spanish print knowledge items measured variance in print knowledge ability that was shared across Spanish and English. All Spanish print knowledge items also measured variance in print knowledge ability that was unique to Spanish. English print knowledge items assessing knowledge of print concepts, letter discrimination, and word discrimination abilities measured variance that was specific to English; however, items measuring knowledge of letter names and letter–sound correspondence did not measure variance that was specific to English.

Omega for the bifactor print knowledge model was estimated with the negative loadings from the full model removed to compute estimates of variance accounted for by the bifactor model. Omega statistics for the print knowledge model are shown in the rightmost columns of Table 4. Omega statistics indicated that approximately 79% of variance in total print knowledge scores was due to the General factor, approximately 17% was due to the Spanish factor, and approximately 3% was due to the English factor. Additionally, 41% of variance in scores on Spanish print knowledge items was due to the General factor, and 59% was due to the Spanish factor. In contrast, 91% of variance in scores on English print knowledge items was due to the General factor, and 9% was due to the English factor.

Oral Language

Model fit statistics for expressive vocabulary are reported in Table 2. For expressive vocabulary, the two-factor model provided significantly better fit to the data than did the one-factor model. The correlation between the English and Spanish Expressive Vocabulary factors was negative and statistically significant ($r = -.12$, $p < .05$). Detailed results of the two-factor model are reported in the left-most columns of Table 7. All Spanish and English expressive vocabulary items loaded on their respective factors. A bifactor model of expressive vocabulary did not converge, indicating that English and Spanish expressive vocabulary items measured only variance unique to the language of the item.

Model fit statistics for definitional vocabulary are reported in the lowest panels of Table 2. The two-factor model provided significantly better fit to the data than did the one-factor model. Detailed results of the two-factor model are reported in the right-most columns of Table 7. The correlation between the Spanish and English Definitional Vocabulary factors was not statistically significant ($r = -.02$, $p = .72$). All Spanish and English items loaded on their respective factors. A bifactor model of definitional vocabulary did not converge, indicating that English and Spanish expressive vocabulary items measured only variance unique to the language of the item.

Cross-Construct Relations

Factor scores estimated from best fitting models were used to examine the relations between the general and specific early literacy factors across constructs. Results of correlational analyses are reported in Table 8. For print knowledge, the orthogonality constraints imposed in the bifactor models were preserved in the factor scores; however, this was not the case for blending and elision factor scores. Within-Spanish factor correlations were positive and statistically significant. Within-English factor correlations were

Table 3
Standardized Factor Loadings From the Bifactor Blending Model With Two Specific Factors (Spanish and English Blending)

Spanish blending			English blending		
Item (and answer) types	Specific	General	Item (and answer) types	Specific	General
1. W (MC)	.45	.40	1. W (MC)	.49	.34
2. W (MC)	.54	.67	2. W (MC)	.47	.52
3. W (MC)	.41	.44	3. W (MC)	.27	.23
4. W (MC)	.42	.54	4. P (MC)	.55	.44
5. W (MC)	.41	.38	5. P (MC)	.54	.55
6. W (MC)	.37	.75	6. P (MC)	.55	.43
7. W (FR)	.78	.35	7. W (FR)	.85	.11†
8. W (FR)	.77	.39	8. W (FR)	.83	.15†
9. W (FR)	.83	.37	9. W (FR)	.82	.05†
10. W (FR)	.82	.33	10. P (FR)	.79	.16*
11. W (FR)	.83	.34	11. P (FR)	.73	.24**
12. W (FR)	.82	.37	12. P (FR)	.86	.21**
13. S (MC)	.56	.69	13. P (FR)	.80	.23*
14. S (MC)	.59	.67	14. P (FR)	.80	.21**
15. S (MC)	.45	.60	15. P (FR)	.85	.18*
16. S (MC)	.53	.66			
17. S (MC)	.42	.35			
18. P (MC)	.40	.58			
19. S (MC)	.52	.52			
20. S (MC)	.53	.62			
21. S (MC)	.57	.64			
22. S (MC)	.64	.59			
23. S (FR)	.88	.25			
24. S (FR)	.90	.24			
25. S (FR)	.90	.24			
26. S (FR)	.88	.28			
27. S (FR)	.89	.23			
28. S (FR)	.90	.25			
29. S (FR)	.91	.23			
30. S (FR)	.92	.20**			
31. S (FR)	.94	.22			
32. S (FR)	.91	.20**			

Note. Correlations between factors are constrained to zero for the estimation of bifactor models. All factor loadings are statistically significant at $p < .001$ unless otherwise noted. W = word; S = syllable; P = phoneme; MC = multiple choice; FR = free response.
$^{\dagger}p > .10$. $^{*}p < .05$. $^{**}p < .01$.

statistically significant for phonological awareness and vocabulary knowledge; however, the Specific English Print Knowledge factor was not consistently related to other Specific English Early Literacy factors. The General Phonological Awareness and Print Knowledge factors were significantly and positively related to each other and to most other factors, with the exception of the Specific English Print Knowledge factor. In general, there were not strong cross-language relations between the Specific English and Spanish factors.

Discussion

The purpose of this study was to evaluate the relevance of the common underlying proficiency model for LM preschooler's Spanish and English early literacy skills. Results indicated that there was a common underlying proficiency for children's code-related but not language-related skills. Specifically, items assessing English and Spanish phonological awareness and print knowledge measured variance that was shared across languages as well as variance specific to English or Spanish. In contrast, items assessing English and Spanish oral language measured only variance unique to the language of the item. Cross-construct correlations indicated that skills unique to Spanish were related to each other and skills unique to English were related to each other (with the exception of English print knowledge). Language-independent phonological awareness and print knowledge abilities were related to each other and to children's early literacy abilities that were unique to Spanish and English. Taken together, these findings indicate that Spanish-speaking LM preschoolers have a common underlying proficiency for phonological awareness and for print knowledge but not for oral language. Evidence for the common underlying proficiency suggests that children's code-related skills can be more easily transferred across languages than can language skills. However, evidence of a common underlying proficiency is not necessarily evidence of cross-language transfer. Further research is needed to better understand the conditions under which LM children can utilize a common underlying proficiency to transfer knowledge across languages.

Code-Related Skills

Prior studies examining the interdependence of L1 and L2 phonological awareness skills have indicated that LM children's phonological awareness skills are significantly related across languages (e.g., Branum-Martin et al., 2006). Significant cross-language correlations of phonological awareness are often interpreted as evidence that children transferred knowledge across languages (e.g., Durgunoğlu et al., 1993). However, it is possible that cross-language correlations are due to other factors, such as common language-learning environments for L1 and L2. Alternatively, cross-language correlations could be indicative of the presence of a common underlying proficiency (Cummins, 1981), which is one mechanism through which cross-language transfer could occur. This is the first study to date to evaluate empirically whether there is a common underlying proficiency for phonological awareness.

Table 4
Omega Values for Bifactor Models of Early Literacy

Variable	Blending	Elision	Print
Omega total	.98	.97	.98
Omega hierarchical			
General factor	.35	.44	.78
Spanish factor	.52	.44	.17
English factor	.12	.09	.03
Omega (Spanish items)	.99	.97	.97
Omega subscale (Spanish items)			
General factor	.28	.43	.39
Spanish factor	.71	.54	.58
Omega (English items)	.95	.89	.98
Omega subscale (English items)			
General factor	.13	.05	.89
English factor	.82	.84	.09

Note. Dividing omega hierarchical by omega total yields the percentage of variance in the total test score attributable to each factor. For each subset of items (i.e., Spanish and English), dividing omega subscale by omega yields the percentage of variance in those items attributable to each factor.

Table 5
Standardized Factor Loadings From the Bifactor Elision Model With Two Specific Factors (English and Spanish Elision)

Spanish elision			English elision		
Item (and answer) types	Specific	General	Item (and answer) types	Specific	General
1. W (MC)	.77	−.11*	1. W (MC)	.63	−.08†
2. W (MC)	.66	.00†	2. W (MC)	.54	.08†
3. W (MC)	.79	−.06†	3. W (MC)	.53	.09†
4. W (MC)	.61	−.08†	4. S (MC)	.52	.04†
5. W (MC)	.87	−.01†	5. S (MC)	.53	−.08†
6. W (MC)	.68	.02†	6. S (MC)	.43	.03†
7. W (FR)	.69	.50	7. W (FR)	.72	.34
8. W (FR)	.63	.52	8. W (FR)	.75	.36
9. W (FR)	.66	.54	9. W (FR)	.82	.35
10. W (FR)	.65	.59	10. P (FR)	.75	.29
11. W (FR)	.66	.60	11. P (FR)	.73	.20
12. W (FR)	.62	.60	12. P (FR)	.73	.25
13. S (MC)	.68	.22			
14. S (MC)	.78	.02†			
15. P (MC)	.55	.06†			
16. S (MC)	.70	.12^{+}			
17. S (MC)	.76	.03†			
18. P (MC)	.79	−.02†			
19. S (MC)	.72	.16**			
20. S (MC)	.67	.08†			
21. P (MC)	.45	.30			
22. S (MC)	.67	.10*			
23. S (FR)	.50	.69			
24. S (FR)	.71	.59			
25. S (FR)	.60	.55			
26. S (FR)	.37	.74			
27. S (FR)	.51	.64			
28. S (FR)	.54	.73			
29. S (FR)	.31	.72			
30. S (FR)	.48	.68			
31. S (FR)	.57	.68			
32. P (FR)	.37	.76			

Note. Correlations between factors are set to zero for the estimation of bifactor models. All factor loadings are statistically significant at $p < .001$ unless otherwise noted. W = word; S = syllable; P = phoneme; MC = multiple choice; FR = free response.
$^{\dagger}p > .10$. $^{+}p < .10$. $^{*}p < .05$. $^{**}p < .01$.

Results of this study indicated that items on phonological awareness assessments measured both language-specific and language-independent variance in phonological awareness ability, suggesting that LM preschoolers have a common underlying proficiency for phonological awareness. With the exception of elision items that were multiple choice, Spanish items had stronger loadings on the General Phonological Awareness factors than did English items, indicating that L1 phonological awareness is a better indicator of the common underlying proficiency than is L2 phonological awareness. Similarly, examination of variance accounted for indicated that more variance in Spanish phonological awareness scores than in English phonological awareness scores was due to the General Phonological Awareness factors. One possible explanation for this finding is the discrepancy in exposure to L1 and L2 for LM children. For many LM children in the United States, substantial exposure to English (i.e., L2) does not occur until preschool entry. Language exposure should be associated with increases in vocabulary knowledge that may lead to improved phonological awareness abilities (Walley, Metsala, & Garlock, 2003). Therefore, L1 phonological awareness assessments may better approximate preschool children's underlying capacity for phonological awareness because of increased opportunities for the development of phonological awareness in L1 that come from language exposure. Consistent with this explanation, evidence has indicated that L1 and L2 phonological awareness are related only for children with higher levels of L1 language skills (Atwill, Blanchard, Christie, Gorin, & García, 2010; Goodrich, Lonigan, & Farver, 2014), suggesting that increased language exposure promotes the development of language-independent phonological awareness abilities. Additionally, the finding that the Specific English factor accounted for a smaller amount of variance in total

Table 6
Standardized Factor Loadings From General and Specific Factors From the Bifactor Model of Spanish and English Print Knowledge

Spanish			English		
Item (and answer) types	Specific	General	Item (and answer) types	Specific	General
1. PC	.35	.25	1. PC	.26	.47
2. PC	.42	.44	2. PC	.42	.38
3. PC	.38	.30	3. PC	.37	.32
4. PC	.33	.27	4. PC	.36	.29
5. LD	.50	.48	5. LD	.64	.63
6. LD	.46	.52	6. LD	.63	.66
7. LD	.44	.50	7. LD	.63	.67
8. LD	.28	.33	8. LD	.59	.48
9. WD	.42	.43	9. WD	.75	.61
10. WD	.43	.45	10. WD	.75	.63
11. WD	.46	.49	11. WD	.75	.58
12. WD	.44	.49	12. WD	.75	.62
13. LN (MC)	.42	.61	13. LN (MC)	−.03†	.73
14. LN (MC)	.26	.46	14. LN (MC)	−.04†	.72
15. LN (MC)	.35	.33	15. LN (MC)	−.03†	.76
16. LN (MC)	.37	.35	16. LN (MC)	−.02†	.73
17. LN (MC)	.46	.36	17. LN (MC)	−.04†	.85
18. LN (MC)	.38	.53	18. LN (MC)	−.02†	.78
19. LS (MC)	.32	.59	19. LS (MC)	.05†	.70
20. LS (MC)	.35	.66	20. LS (MC)	−.02†	.76
21. LS (MC)	.30	.63	21. LS (MC)	−.03†	.83
22. LS (MC)	.35	.63	22. LS (MC)	−.01†	.83
23. LN (FR)	.85	.23	23. LN (FR)	−.05†	.69
24. LN (FR)	.84	.19	24. LN (FR)	−.13†	.75
25. LN (FR)	.87	−.05†	25. LN (FR)	−.04†	.73
26. LN (FR)	.82	−.16**	26. LN (FR)	−.15†	.80
27. LN (FR)	.91	.16**	27. LN (FR)	−.12†	.84
28. LN (FR)	.84	.27	28. LN (FR)	−.16†	.80
29. LN (FR)	.93	.13*	29. LN (FR)	−.20^{+}	.87
30. LN (FR)	.89	−.06†	30. LN (FR)	−.15†	.81
31. LN (FR)	.89	.13*	31. LN (FR)	−.08†	.86
32. LN (FR)	.81	.19**	32. LN (FR)	−.10†	.61
33. LS (FR)	.48	.55	33. LS (FR)	−.19†	.85
34. LS (FR)	.41	.60	34. LS (FR)	−.13†	.91
35. LS (FR)	.43	.64	35. LS (FR)	−.16†	.87
36. LS (FR)	.46	.58	36. LS (FR)	−.20†	.87

Note. Correlations between factors are set to zero for the estimation of bifactor models. All factor loadings are statistically significant at $p < .001$ unless otherwise noted. PC = print concepts; LD = letter discrimination; WD = word discrimination; LN = letter–name identification; LS = letter–sound identification; MC = multiple choice; FR = free response.
$^{\dagger}p > .10$. $^{+}p < .10$. $^{*}p < .05$. $^{**}p < .01$.

Table 7
Standardized Factor Loadings for English and Spanish Vocabulary Factors From Two-Factor Models of Expressive and Definitional Vocabulary

Expressive vocabulary			Definitional vocabulary		
Item	Spanish	English	Item	Spanish	English
1	.65	.75	1	.70	.75
2	.62	.72	2	.62	.47
3	.67	.53	3	.65	.63
4	.73	.75	4	.63	.53
5	.65	.63	5	.69	.69
6	.78	.78	6	.72	.58
7	.68	.76	7	.59	.61
8	.63	.79	8	.64	.54
9	.74	.81	9	.66	.43
10	.76	.66	10	.68	.42
11	.76	.89	11	.69	.65
12	.84	.82	12	.69	.59
13	.67	.66	13	.63	.68
14	.76	.75	14	.68	.50
15	.81	.63	15	.67	.57
16	.79	.68	16	.80	.72
17	.84	.76	17	.81	.64
18	.66	.60	18	.67	.51
19	.72	.42	19	.82	.69
20	.90	.87	20	.88	.77
21	.80	.75	21	.75	.51
22	.82	.62	22	.80	.61
23	.88	.66	23	.85	.76
24	.92	.81	24	.71	.48
25	.92	.77	25	.85	.60
26	.80	.53	26	.93	.86
27	.83	.44	27	.92	.85
28	.85	.74	28	.85	.71
29	.77	.60	29	.86	.68
30	.83	.58	30	.81	.63
31	.84	.78	31	.87	.70
32	.62	.11†	32	.95	.77
33	.88	.60	33	.83	.49
34	.70	.58	34	.86	.59
35	.59	.46	35	.77	.50

Note. All factor loadings are significant at $p < .001$ unless otherwise noted.
† $p > .10$.

phonological awareness abilities than did the Specific Spanish factor or the General factor may be an artifact of the amount of exposure to each language. If LM children's Spanish phonological awareness skills are more advanced relative to their English phonological awareness skills, items on Spanish assessments should be better indicators of children's underlying capacity for phonological awareness, which was the pattern of results obtained in this study.

As was the case for phonological awareness, results of this study indicated that LM preschoolers have a common underlying proficiency for print knowledge. Prior research has indicated that children's print knowledge is significantly related across languages (e.g., Anthony et al., 2009). Although there is some language-independent information about print knowledge for which children could have a common underlying proficiency (e.g., the knowledge that letters have names and are associated with sounds), there is also language-specific information about print knowledge (e.g., letter names and letter–sound correspondences differ across languages). Therefore, the extent to which print knowledge is related across languages may be limited by the degree of similarity in letter names and letter–sound correspondence across languages. For example, many letters in English and Spanish have similar names, and several letters correspond to the same sounds in English and Spanish. L1 and L2 print knowledge skills may not be as highly related when the alphabetic system differs across languages (e.g., English–Arabic) or for alphabetic and nonalphabetic languages (e.g., English–Chinese). However, McBride-Chang and Ho (2005) reported that knowledge of letter names in English was significantly correlated concurrently and longitudinally with Chinese character recognition, which may be indicative of transfer of language-independent information that pertains to print knowledge.

In contrast to results for phonological awareness, English print knowledge items had stronger loadings on the General Print Knowledge factor than did Spanish print knowledge items, and the General Print Knowledge factor accounted for a larger amount of variance in scores for English print knowledge items than it did for Spanish print knowledge items. This is likely because knowledge of letter names and letter–sound correspondence is language-specific information that is explicitly taught, whereas phonological awareness is a language-independent ability that may be a developmental consequence of language exposure (Walley et al., 2003). Because LM children in the United States are primarily instructed in English, English print knowledge assessments may be better indicators of children's common underlying proficiency for print knowledge than are Spanish print knowledge assessments. When there is substantial overlap in letter names and letter–sound correspondence across languages (as is the case for English and Spanish), children may be able to apply knowledge gained from L2 print knowledge instruction to their L1, in which letter names and letter–sound correspondence may not have been explicitly taught.

Contrary to hypotheses, the common underlying proficiency model was more relevant for print knowledge than it was for phonological awareness, because the total variance accounted for by the General factor was higher for print knowledge than it was for both blending and elision. This finding is consistent with a stronger cross-language correlation in the two-factor model for print knowledge than in the two-factor models for phonological awareness. Because the primary language of instruction to which many LM children in the United States are exposed is English and there is a large degree of overlap in language-specific aspects of print knowledge across English and Spanish (e.g., letter–sound correspondence), the print knowledge skills of Spanish-speaking LM preschoolers in the United States may be better represented by a common underlying proficiency than are phonological awareness abilities.

Oral Language

In contrast to results for code-related skills, there was no evidence of a common underlying proficiency for LM preschoolers' language skills. Expressive vocabulary knowledge was negatively correlated across languages, and definitional vocabulary knowledge was not correlated across languages. This finding was consistent with prior evidence that LM children's vocabulary knowledge is distributed across their two languages (e.g., Peña et al., 2002) and that vocabulary knowledge is not correlated or is neg-

Table 8
Correlations Between Specific and General Early Literacy Factors

Factor	1	2	3	4	5	6	7	8	9	10	11	12
1. Elision—S	—											
2. Elision—E	**.21****	—										
3. Elision—G	**.20****	**.08***	—									
4. Blend—S	.52***	.13***	.39***	—								
5. Blend—E	.12**	.43***	.20***	**.17*****	—							
6. Blend—G	.56***	.22***	.13***	**.18*****	**.08***	—						
7. Print—S	.37***	.06^{+}	.32***	.39***	.04	.29***	—					
8. Print—E	.00	.12*	.07*	.04	.07^{+}	.05	**.10****	—				
9. Print—G	.36***	.33***	.27***	.29***	.43***	.38***	**.05**	**.02**	—			
10. EV—S	.26***	.03	.20***	.29***	.08*	.20***	.32***	−.01	.14***	—		
11. EV—E	.10**	.19***	.08*	−.02	.19***	.16***	−.05	.00	.26***	−.15***	—	
12. DV—S	.26***	.03	.18***	.29***	.08*	.21***	.32***	.01	.15***	**.95*****	−.11**	—
13. DV—E	.12**	.20***	.10**	.02	.20***	.17***	.01	.00	.27***	−.07^{+}	**.85*****	−.02

Note. Correlations shown in boldface are within-construct correlations. S = Spanish; E = English; G = General; EV = Expressive Vocabulary; DV = Definitional Vocabulary.
$^{+}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

atively correlated across languages (e.g., Bialystok et al., 2005; Goodrich et al., 2016). Words in a language are arbitrarily associated with their underlying concepts, and words used to describe the same object in two different languages are often remarkably different (with the exception of cognates). Therefore, unless words across English and Spanish are cognates, there is little to no information concerning the form of a word in L1 that could be applied to L2 to assist in the acquisition of its translation equivalent. However, the lack of a common underlying proficiency does not entirely rule out all types of transfer of language skills, as prior evidence has suggested that children transfer some word-specific information across languages (Goodrich et al., 2016).

Relations Between Language-Independent and Language-Specific Aspects of Early Literacy

Consistent with hypotheses, the common underlying proficiencies for phonological awareness and print knowledge were significantly related to language-specific early literacy abilities across constructs (e.g., general phonological awareness had similar relations with English and Spanish expressive vocabulary). This finding suggests that the common underlying proficiencies for code-related skills represent a general language-learning capacity (Castilla et al., 2009) rather than construct-specific abilities. Specific English and Spanish early literacy abilities were generally related to each other within but not across languages (e.g., specific Spanish phonological awareness was related to Spanish expressive vocabulary but not English expressive vocabulary). One unexpected finding was that English-specific aspects of print knowledge were not consistently related to other language-independent or language-specific early literacy abilities. It is possible that this finding is a result of overextraction of variance in English print knowledge items by the General Print Knowledge factor, resulting in weaker and potentially less reliable loadings on the Specific English Print Knowledge factor. Consistent with this explanation, loadings on the Specific English Print Knowledge factor were negative and nonsignificant for items assessing knowledge of letter names and letter–sound correspondence.

Implications

The finding that LM children's code-related, but not language-related, early literacy skills are represented by a common underlying proficiency has implications for researchers and practitioners. It is possible that examining the longitudinal relations between language-specific and language-independent aspects of early literacy and LM children's conventional reading skills may reveal patterns of relations between L1 and L2 different from those that have been highlighted in prior research. For example, although prior studies have reported that L1 phonological awareness skills predict L2 reading outcomes (e.g., Sparks, Patton, Ganschow, Humbach, & Javorksy, 2008), it is possible that only variance in L1 phonological awareness scores that is common to both L1 and L2 accounts for subsequent L2 reading outcomes. Future research should examine the longitudinal predictive validity of language-specific and language-independent aspects of early literacy. Additionally, the presence of language-independent early literacy skills suggests that instruction in L1 will improve code-related skills in both L1 and L2. Therefore, for code-related skills like print knowledge and phonological awareness, evidence-based instruction in either language should provide LM preschoolers with the foundational skills that they need to succeed when formal reading instruction begins. However, this is not necessarily the case for children's early language skills. The results of this study suggest that beneficial effects of language exposure and instruction will be seen in only the language of instruction. This pattern of results may explain the typical finding that LM children have code-related skills that are approximately equivalent to those of their monolingual peers, despite significantly lower language skills. It is important that the results of this study be interpreted with caution, because more research is needed to understand completely how L1 and L2 language skills develop and how educators of LM children can maximize children's academic outcomes. For example, some studies have suggested that among older LM children L1 vocabulary knowledge is uniquely predictive of L2 reading

comprehension above and beyond the effects of L2 reading (e.g., Proctor, August, Carlo, & Snow, 2006), a finding that is in contrast to the pattern of results obtained in this study.

Limitations and Future Directions

Although this study had numerous strengths (e.g., sample size, diversity of LM children within this sample), it also had several limitations. First, this study did not control for other factors related to language and literacy acquisition, such as children's overall cognitive ability. It is possible that the General factor extracted in bifactor models does not represent a common underlying proficiency for a skill but rather overall level of cognitive ability. Additionally, the sample in this study was intended to represent an at-risk, low-income sample of preschoolers. Future studies should evaluate the relevance of the common underlying proficiency model for LM children with a wide range of skills from various demographic backgrounds. Furthermore, item-level data was used in this study, limiting the analytic options available. For example, many parameters were estimated in bifactor models, and more complex models (e.g., multilevel measurement models) could not be estimated because the number of parameters exceeded the number of cluster units in the data. Future research should attempt to replicate the findings of this study using scale-level data that would allow for the evaluation of more complex models (e.g., multilevel bifactor models) that could not be estimated in this study because item-level data were used. Finally, it is possible that a different pattern of results would be obtained with older LM children. For example, the oral language assessments used with preschool children in this study included mostly concrete words, and it is possible that knowledge of words that correspond to more abstract concepts is more easily shared across languages. Future research should examine whether the same pattern of results emerges for LM children across development.

Conclusions

Results of this study indicated that LM preschoolers' code-related early literacy skills were best characterized by a common underlying proficiency as well as specific Spanish and English skills. In contrast, no evidence for a common underlying proficiency for oral language skills emerged, indicating that LM preschoolers' oral language skills were best characterized by specific Spanish and English skills. Evidence in support of a common underlying proficiency for early literacy skills has suggested that cross-language transfer of literacy-related skills is possible, because children should be able to apply language-independent knowledge gained from L1 when learning L2, or vice versa. Future research is needed to determine the relative predictive validity of language-independent and language-specific aspects of early literacy and to better understand how children's language and literacy environments foster development of a common underlying proficiency and language-specific aspects of early literacy.

References

Anthony, J. L., Solari, E. J., Williams, J. M., Schoger, K. D., Zhang, Z., Branum-Martin, L., & Francis, D. J. (2009). Development of bilingual phonological awareness in Spanish-speaking English language learners: The roles of vocabulary, letter knowledge, and prior phonological awareness. *Scientific Studies of Reading, 13,* 535–564. http://dx.doi.org/10.1080/10888430903034770

Asparouhov, T., & Muthén, B. (2013). *Computing the strictly positive Satorra-Bentler Chi-Square Test* (Mplus Web Note No.12). Retrieved from http://www.statmodel.com/examples/webnote.shtml

Atwill, K., Blanchard, J., Christie, J., Gorin, J. S., & García, H. S. (2010). English-language learners: Implications of limited vocabulary for cross-language transfer of phonemic awareness with kindergarteners. *Journal of Hispanic Higher Education, 9,* 104–129. http://dx.doi.org/10.1177/1538192708330431

August, D., Shanahan, T., & Escamilla, K. (2009). English language learners: Developing literacy in second-language learners—Report of the National Literacy Panel on language-minority children and youth. *Journal of Literacy Research, 41,* 432–452. http://dx.doi.org/10.1080/10862960903340165

Bialystok, E., Luk, G., & Kwan, E. (2005). Bilingualism, biliteracy, and learning to read: Interactions among languages and writing systems. *Scientific Studies of Reading, 9,* 43–61. http://dx.doi.org/10.1207/s1532799xssr0901_4

Branum-Martin, L., Mehta, P. D., Fletcher, J. M., Carlson, C. D., Ortiz, A., Carlo, M., & Francis, D. J. (2006). Bilingual phonological awareness: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classrooms. *Journal of Educational Psychology, 98,* 170–181. http://dx.doi.org/10.1037/0022-0663.98.1.170

Byrne, B., & Fielding-Barnsley, R. (1995). Evaluation of a program to teach phonemic awareness to young children: A 2- and 3-year follow-up and a new preschool trial. *Journal of Educational Psychology, 87,* 488–503. http://dx.doi.org/10.1037/0022-0663.87.3.488

Castilla, A. P., Restrepo, M. A., & Perez-Leroux, A. T. (2009). Individual differences and language interdependence: A study of sequential bilingual development in Spanish-English preschool children. *International Journal of Bilingual Education and Bilingualism, 12,* 565–580. http://dx.doi.org/10.1080/13670050802357795

Cheung, A. C. K., & Slavin, R. E. (2012). Effective reading programs for Spanish-dominant English language learners (ELLs) in the elementary grades: A synthesis of research. *Review of Educational Research, 51,* 879–912.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49,* 222–251. http://dx.doi.org/10.3102/00346543049002222

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education. (Ed.), *Schooling and language minority students: A theoretical framework* (pp. 3–49). Los Angeles, CA: Evaluation, Dissemination and Assessment Center, California State University, Los Angeles.

Cummins, J. (2008). Teaching for transfer: Challenging the two solitudes assumption in bilingual education. In J. Cummins & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 5. Bilingual education* (2nd ed., pp. 65–75). New York, NY: Springer Science + Business Media.

Durgunoğlu, A. Y., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. *Journal of Educational Psychology, 85,* 453–465. http://dx.doi.org/10.1037/0022-0663.85.3.453

Farver, J. A., Lonigan, C. J., & Eppe, S. (2009). Effective early literacy skill development for young Spanish-speaking English language learners: An experimental study of two methods. *Child Development, 80,* 703–719. http://dx.doi.org/10.1111/j.1467-8624.2009.01292.x

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment, 30,* 130–139. http://dx.doi.org/10.1027/1015-5759/a000181

Goodrich, J. M., Lonigan, C. J., & Farver, J. M. (2013). Do early literacy skills in children's first language promote development of skills in their second language? An experimental evaluation of transfer. *Journal of Educational Psychology, 105,* 414–426. http://dx.doi.org/10.1037/a0031780

Goodrich, J. M., Lonigan, C. J., & Farver, J. M. (2014). Children's expressive language skills and their impact on the relation between first- and second-language phonological awareness skills. *Scientific Studies of Reading, 18,* 114–129. http://dx.doi.org/10.1080/10888438.2013.819355

Goodrich, J. M., Lonigan, C. J., Kleuver, C. G., & Farver, J. M. (2016). Development and transfer of vocabulary knowledge in Spanish-speaking language minority preschool children. *Journal of Child Language, 43,* 969–992. http://dx.doi.org/10.1017/S030500091500032X

Gottardo, A., & Mueller, J. (2009). Are first- and second-language factors related in predicting second-language reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology, 101,* 330–344. http://dx.doi.org/10.1037/a0014320

Hemphill, F. C., Vanneman, A., & Rahman, T. (2011). *Achievement gaps: How Hispanic and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress: Statistical analysis report* (NCES 2011-459). Washington, DC: National Center for Education Statistics.

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology, 49,* 4–14. http://dx.doi.org/10.1037/a0027238

Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., & Snowling, M. J. (2012). The causal role of phoneme awareness and letter–sound knowledge in learning to read: Combining intervention studies with mediation analyses. *Psychological Science, 23,* 572–577. http://dx.doi.org/10.1177/0956797611435921

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. http://dx.doi.org/10.1080/01621459.1995.10476572

Kieffer, M. J., & Vukovic, R. K. (2013). Growth in reading-related skills of language minority learners and their classmates: More evidence for early identification and intervention. *Reading and Writing, 26,* 1159–1194. http://dx.doi.org/10.1007/s11145-012-9410-7

Lindsey, K. A., Manis, F. R., & Bailey, C. E. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. *Journal of Educational Psychology, 95,* 482–494. http://dx.doi.org/10.1037/0022-0663.95.3.482

Lonigan, C. J. (2012). *Spanish preschool early literacy assessment.* Tallahassee, FL: Author.

Lonigan, C. J., Farver, J. M., & Eppe, S. (2002). *Preschool comprehensive test of phonological and print processing: Spanish version.* Tallahassee, FL: Author.

Lonigan, C. J., Farver, J. M., Nakamoto, J., & Eppe, S. (2013). Developmental trajectories of preschool early literacy skills: A comparison of language-minority and monolingual-English children. *Developmental Psychology, 49,* 1943–1957. http://dx.doi.org/10.1037/a0031408

Lonigan, C. J., Schatschneider, C., & Westberg, L. (2008). Identification of children's skills and abilities linked to later outcomes in reading, writing, and spelling. In *Developing early literacy: Report of the National Early Literacy Panel* (pp. 55–106). Washington, DC: National Institute for Literacy.

Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2007). *Test of preschool early literacy.* Austin, TX: Pro-Ed.

Manis, F. R., Lindsey, K. A., & Bailey, C. E. (2004). Development of reading in grades K-2 in Spanish-speaking English-language learners. *Learning Disabilities Research & Practice, 19,* 214–224. http://dx.doi.org/10.1111/j.1540-5826.2004.00107.x

McBride-Chang, C., & Ho, C. S.-H. (2005). Predictors of beginning reading in Chinese and English: A 2-year longitudinal study of Chinese kindergarteners. *Scientific Studies of Reading, 9,* 117–144. http://dx.doi.org/10.1207/s1532799xssr0902_2

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading, 34,* 114–135. http://dx.doi.org/10.1111/j.1467-9817.2010.01477.x

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Peña, E. D., Bedore, L. M., & Zlatic-Giunta, R. (2002). Category-generation performance of bilingual children: The influence of condition, category, and language. *Journal of Speech, Language, and Hearing Research, 45,* 938–947. http://dx.doi.org/10.1044/1092-4388(2002/076)

Proctor, C. P., August, D., Carlo, M. S., & Snow, C. (2006). The intriguing role of Spanish language vocabulary knowledge in predicting English reading comprehension. *Journal of Educational Psychology, 98,* 159–169. http://dx.doi.org/10.1037/0022-0663.98.1.159

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95,* 129–140. http://dx.doi.org/10.1080/00223891.2012.725437

Sparks, R. L., Patton, J., Ganschow, L., Humbach, N., & Javorsky, J. (2008). Early first-language reading and spelling skills predict later second-language reading and spelling skills. *Journal of Educational Psychology, 100,* 162–174. http://dx.doi.org/10.1037/0022-0663.100.1.162

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38,* 934–947. http://dx.doi.org/10.1037/0012-1649.38.6.934

Walley, A. C., Metsala, J. L., & Garlock, V. M. (2003). Spoken vocabulary growth: Its role in the development of phoneme awareness and early reading ability. *Reading and Writing, 16,* 5–20. http://dx.doi.org/10.1023/A:1021789804977

Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development, 69,* 848–872. http://dx.doi.org/10.1111/j.1467-8624.1998.tb06247.x

Received June 20, 2016
Revision received November 28, 2016
Accepted December 1, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 6, 794–811

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000165

Differential Effects of the Classroom on African American and Non–African American's Mathematics Achievement

Katerina Schenke
University of California, Los Angeles

Tutrang Nguyen and Tyler W. Watts
University of California, Irvine

Julie Sarama and Douglas H. Clements
University of Denver

We examined whether African American students differentially responded to dimensions of the observed classroom-learning environment compared with non–African American students. Further, we examined whether these dimensions of the classroom mediated treatment effects of a preschool mathematics intervention targeted at students from low-income families. Three observed dimensions of the classroom (teacher expectations and developmental appropriateness; teacher confidence and enthusiasm; and support for mathematical discourse) were evaluated in a sample of 1,238 preschool students in 101 classrooms. Using multigroup multilevel mediation where African American students were compared with non–African American students, we found that teachers in the intervention condition had higher ratings on the observed dimensions of the classroom compared with teachers in the control condition. Further, ratings on teacher expectations and developmental appropriateness had larger associations with the achievement of African American students than for non–African Americans. Findings suggest that students within the same classroom may react differently to that learning environment and that classroom learning environments could be structured in ways that are beneficial for students who need the most support.

Educational Impact And Implications Statement

Data from this study come from a randomized control trial of a preschool mathematics intervention. We used multigroup multilevel structural equation modeling to test whether certain instructional practices explained why African American students had differential gains in mathematics achievement as compared with non-African American students. This study suggests that, on average, African American students benefit differently from certain instructional practices (e.g., teacher expectations and developmental appropriateness) more than non-African American students. Implications suggest that content-specific interventions could be designed to include changing teachers' beliefs about students' abilities.

Keywords: achievement gap, classroom observations, mathematics learning, multilevel structural equation modeling

Supplemental materials: http://dx.doi.org/10.1037/edu0000165.supp

Identifying ways to promote student learning, especially for marginalized groups, is an important goal of educational research. On average, the mathematics achievement gap between low-income minority students and high-income White students is close to two thirds of a standard deviation at the start of kindergarten (Duncan & Magnuson, 2005; Loeb & Bassok, 2008; Reardon & Robinson, 2008). This has led both researchers and educational advocacy groups to call for high-quality mathematics instruction prior to school-entry, as such efforts could reduce achievement gaps at the beginning, and throughout, K–12 schooling (National

This article was published Online First March 23, 2017.

Katerina Schenke, Graduate School of Education & Information Studies, University of California, Los Angeles; Tutrang Nguyen and Tyler W. Watts, School of Education, University of California, Irvine; Julie Sarama and Douglas H. Clements, Morgridge College of Education, University of Denver.

This research was supported by the Institute of Education Sciences, U.S. Department of Education through Grants R305A120813, R305K05157, and R305A110188 and the NICHD-supported Irvine Network on Interventions in Development (HD065704 P01). The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education. The authors express appreciation to the school districts, teachers, and children who participated in this research. We also thank Greg Duncan and Sandra Graham for comments on earlier drafts of this article and Erik Ruzek and Kevin Schaaf for their consultation on the analysis.

Correspondence concerning this article should be addressed to Katerina Schenke, Graduate School of Education & Information Studies, University of California, Los Angeles, CA 90095-1521. E-mail: kschenke@ucla.edu

Council of Teachers of Mathematics (NCTM), 2000; National Mathematics Advisory Panel, 2008; National Research Council, 2009).

The most critical feature of a high-quality educational environment is a knowledgeable and responsive adult (Darling-Hammond, 1997; Ferguson, 1991; National Research Council, 2009; Sarama, & DiBiase, 2004; Schoen, Cebulla, Finn, & Fi, 2003). This may be especially true for classroom-based interventions targeted at closing racial achievement gaps, as certain instructional practices may be more important for some racial/ethnic groups than others (e.g., Bodovski & Farkas, 2007; Sonnenschein & Galindo, 2015; Wenglinsky, 2004). However, understanding which specific elements of the classroom, and whether students within the same classroom respond differently to the same classroom practices, has been largely unexplored. This is an important endeavor as promoting the achievement for marginalized groups, notably African American children, remains scarce (Stinson, 2006).

Data for this study come from a randomized control trial of an intervention that was shown to have a larger effect—almost double in magnitude—on the average mathematics achievement of African American students than of non–African American students (Clements, Sarama, Spitler, Lange, & Wolfe, 2011). Analyses suggest that classroom quality—measured broadly—mediated this effect (Clements et al., 2011). However, it is unknown what specific processes within the classroom led to this larger treatment effect. The goal of the current paper is to explore whether specific classroom practices could explain why African American students, on average, fared better than non–African American students at the end of the intervention. The results of our analyses have broader implications for designing classroom-learning environments to support the learning of marginalized students.

Inequality in Early Learning Experiences

Children with low early academic achievement and those who encounter early learning problems face continuing negative consequences that accumulate over time (Alexander, Entwisle, & Horsey, 1997; Brooks-Gunn & Duncan, 1997; Duncan, Brooks-Gunn, & Klebanov, 1994; Fryer & Levitt, 2006; Huston & Bentley, 2010). This is particularly salient for low-income, minority children who typically begin school with fewer academic skills than their middle- to high-income peers (Duncan & Magnuson, 2005; Lee & Burkam, 2002; Loeb & Bassok, 2008; Reardon & Robinson, 2008). Differences between the educational experiences of low-income and high-income students have been well documented in the literature (Aikens & Barbarin, 2008; Kozol, 1991; Oakes, 1990). However, racial/ethnic differences in students' access to high quality learning environments still exist even when socioeconomic factors are controlled (Lubienski, 2002; Oakes, 1990). In general, race and socioeconomic status are highly correlated, making the study of the unique contributions of race and socioeconomic status difficult to untangle.

Even when minority students have access to the same schools and classroom environments, inequalities exist in how students of different minority groups experience the classroom environment. For example, teachers have been shown to have differential expectations for students of different racial/ethnic groups even when the previous achievement of these students was equivalent (McKown & Weinstein, 2008). Further, when looking within the same functional learning environment, the effects that certain instructional practices have on students have been found to differ according to racial/ethnic group. For example, African Americans on average, compared with students of other races/ethnicities, have been found to differentially benefit from certain instructional practices such as collaborative problem solving (Lubienski, 2006), an emphasis on specific mathematics content (Bodovski & Farkas, 2007; Sonnenschein & Galindo, 2015; Wenglinsky, 2004), and more opportunities to learn higher-level mathematics such as reasoning and problem solving (Battey, 2013; Bodovski & Farkas, 2007; Boaler, 1998; Ladson-Billings, 1997; Lubienski, 2002; Silver & Stein, 1996).

Unfortunately, we still lack a thorough understanding of why African American students may differentially benefit from certain instructional practices compared with students from other racial/ethnic groups. Some scholars suggest that African American students are oriented toward different learning styles (Berry, 2003; Ladson-Billings, 1997) and that these different styles of learning should be considered during teaching. García Coll and colleagues (1996) proposed an integrative model for studying child development and suggested that elements such as the learning environment and culture influence students' psychological experiences. Even when children are in the same environment, their interpretations of certain aspects of the environment may vary according the child's previous cultural experiences, values, and goals. Differences in students' classroom perceptions have been observed in empirical work that has shown that African American students' attitudes toward learning may differ from those of White students (Lubienski, 2002; Strutchens & Silver, 2000). This broad theory proposed by García Coll and colleagues (1996) outlines possible mechanisms for understanding the experiences of minority students, but no theory to date specifically addresses students' differential responsiveness to classroom practices.

Dimensions of the Classroom Learning Environment

Early childhood environments have been described along a variety of dimensions (see Stipek & Byler, 2004 for a review), such as teacher–student interactions (Abbott-Shim, Lambert, & McCarty, 2000; Harms, Clifford, & Cryer, 1998; Pianta & Hamre, 2009; Stipek & Byler, 2004; NICHD ECCRN, 1996), quality of instruction (Abbott-Shim, Lambert, & McCarty, 2000; Pianta & Hamre, 2009), types of activities done in the classroom (Harms et al., 1998), behavior management (Abbott-Shim, Lambert, & McCarty, 2000; Pianta & Hamre, 2009; Stipek & Byler, 2004), and others. Empirically, observational measures are typically used to examine associations between these dimensions and children's developmental outcomes (e.g., Bryant, Burchinal, Lau, & Sparling, 1994; Peisner-Feinberg & Burchinal, 1997). In the present study, we described the classroom environment as (a) teachers' expectations and the developmental appropriateness of their instruction (developmental appropriateness as operationalized by the expectations teachers have of what preschool students are capable of learning), (b) teachers' confidence and enthusiasm in their teaching, and (c) teachers' support for mathematical discourse.

Teacher Expectations

The expectations teachers have of their students may have an effect on students' learning (e.g., Brophy, 1986; Gill & Reynolds, 1999; Rosenthal & Jacobson, 1968) and are related to high quality and equitable instruction (Askew, Brown, Rhodes, Wiliam, & Johnson, 1997; Clarke, Frazer, DiMartino, Fisher, & Smith, 2003; Clements & Sarama, 2007, 2008; NCTM, 2000). These expectations influence whether teachers decide to provide or constrain the opportunities they give their students (Brophy & Good, 1970). For example, a teacher who has high expectations for a student may give that student more opportunities to answer questions during class or further press the student to explain their thinking. In contrast, a teacher who has low expectations for a student may not give that student enough time to respond to a question or may not even call on that student in the first place, thereby making the student miss a key learning opportunity. Many teachers hold low expectations for students whom they believe have low ability or achievement (Brophy & Good, 1970), and these low expectations are often related to the ethnicity of the student (Dusek & Joseph, 1983; Jussim, Eccles, & Madon, 1996; Madon, Jussim, Keiper, Eccles, Smith, & Palumbo, 1998; McKown & Weinstein, 2008). Indeed, teachers who hold particularly low academic expectations for African American students spend more time on behavioral corrections than content instruction (Gill & Reynolds, 1999; Jussim, 1989; Jussim et al., 1996; Kuklinski & Weinstein, 2001; Madon et al., 1998; Raudenbush, 1984; Steele, 1997).

Teacher expectations are often measured at the student level, whereby teachers are asked to complete surveys regarding specific students in their classroom (e.g., McKown & Weinstein, 2008). As such, this method does not readily allow investigation of between-teacher levels of expectations where comparisons across teachers' expectations for their students can be made. One exception is Rubie-Davies (2007), who created categories (low, average, and high) of teachers from their individual ratings of students such that teachers' expectations of students are nested within the teacher. However, this method may mask variation between teachers' expectations. If one imagines teachers' ratings of expectations for each student spanning a normal distribution, a teacher's ratings will vary about the mean. However, once these ratings are aggregated to the teacher-level, thereby calculating a teacher-level rating of expectations, differences *between* teacher's ratings will appear small. Furthermore, understanding differences between teachers' expectations of their students leads to questions about whether interventions can be designed to change these mean-level expectations. This topic has not been exhaustively explored through the implementation of interventions (for exceptions see Proctor, 1984; Rubie-Davies, Peterson, Sibley, & Rosenthal, 2015; Weinstein et al., 1991).

As an exception, Weinstein and colleagues (1991) designed an extensive intervention to change the school and classroom climate to improve low achieving ninth grade students' academic performance and behavior. Whereas one component of this intervention was to raise teachers' expectations of their students, other elements of the classroom-learning environment were also targeted, such as adapting readings from the honors English courses and using heterogeneous grouping strategies, thereby holding low achieving students to higher standards. In another example, Rubie-Davies and colleagues (2015) designed an intervention providing teachers with professional development (PD) aimed at increasing the expectations teachers had of their students. Even though the intervention was described as one focusing on teacher expectations, the PD also focused on other elements of the classroom climate such as increasing student motivation, providing useful feedback, and providing opportunities for promoting student autonomy. Therefore, the effects of the intervention on students' academic achievement could be attributed to all of these classroom dimensions. Whereas previous studies (e.g., Rubie-Davies et al., 2015; Weinstein et al., 1991) hypothesized that part of their observed academic gains were attributable to changes in teachers' classroom behaviors, the present study is the first to empirically test these hypotheses by directly measuring elements of the teacher's classroom practices and by conducting mediational analyses.

Understanding what scholars and practitioners consider as developmentally appropriate in the early years has been questioned. For example, some argue that intensive content-specific instruction in preschool can provide the basic foundational skills that can help prepare children for the academic nature of kindergarten (Ginsburg, Inoue, & Seo, 1999; Seo & Ginsburg, 2004). Others argue that subjecting children to harsh forms of instruction and imposing material they are not ready to learn can be detrimental, forcing young children to engage in developmentally inappropriate forms of drills and practices in mathematics (Bishop-Josef & Zigler, 2011). Young children can indeed engage in various types of mathematical activity if provided with the appropriate opportunities to do so; however, teachers often fail to provide such opportunities to their students or may believe that such practices are developmentally inappropriate (Dunn & Kontos, 1997; Hitz & Wright, 1988). A central premise of the current study is that the developmental needs of students are met when students are held to high expectations, that is, when teachers believe students can engage in higher-level thinking. Preschool children—especially low-income, minority children—have the potential to learn and acquire math skills and concepts (Clements, Baroody, & Sarama, 2013; Ginsburg, Lee, & Boyd, 2008). If teachers and those who work with teachers underestimate what children already know and can learn, they will not present appropriate and challenging mathematics activities.

Teachers' Confidence and Enthusiasm

The beliefs teachers have about their teaching influence their instructional practices (Pajares, 1992; Stipek, 1998; Thompson, 1984) and, in turn, their students' achievement (Evertson, Emmer, & Brophy, 1980). Descriptions of effective mathematics classrooms likely include a teacher who sparks joy and excitement in her students. The ability for a teacher to instruct in a confident and enthusiastic manner is thought to instill confidence and enthusiasm in students and therefore promote learning and achievement. Though confidence and enthusiasm are thought to be important dimensions of the classroom climate, these attributes are rarely measured in quantitative studies of instructional practices.

One exception is an early study by Evertson and colleagues (1980) who conducted observations of less- and more-effective teachers as measured by student achievement on a state standardized test, and found differences in observed enthusiasm for teaching and confidence between the two groups of teachers. Another exception is a study by Stipek and colleagues (2001), who measured teacher confidence and enthusiasm through observations and teacher self-report. They found that teacher's self-reported confidence in teaching mathematics was correlated with the students' classroom average self-confidence in mathematics. Further, they also reported that teachers' enjoyment of mathematics correlated with students' enjoyment of mathematics. Although theory would support the association between teachers' confidence and enthusiasm and student achievement (Stipek, Givvin, Salmon, & MacGyvers, 2001), the authors are not aware of any studies directly testing this association. Evertson and colleagues (1980) only provide weak evidence of such an association as these practices were not directly correlated with student achievement, but rather were only investigated in the context measuring a subsample of teachers chosen by the researchers to examine differences between more effective and less effective teachers.

Support for Mathematical Discourse

Support for mathematical discourse has been recognized as an important practice for fostering students' mathematical learning (Hiebert, & Grouws, 2007; Huttenlocher, Vasilyeva, Waterfall, Vevea, & Hedges, 2007; NCTM, 1991). For example, allowing students the opportunity to explain their thinking, elaborate on concepts, and generate mathematical talk, have been identified as high-level mathematical activities (Henningsen & Stein, 1997; NCTM, 1991) and these practices have effects on students' mathematical knowledge development (Walshaw & Anthony, 2008). Such practices are thought to help students increase their understanding of mathematics by helping students better internalize mathematical content and by allowing other students to learn from hearing student-generated explanations (Chi, 2000). Additionally, when students generate their own mathematical explanations, teachers are able to tailor their instruction to address inaccuracies in explanations or misconceptions (Franke, Fennema, & Carpenter, 1997). However, many teachers do not talk to their students about mathematics, even when the student initiates the mathematical talk. In one study, when students made a mathematical utterance, their teachers ignored it 60% of the time and only responded mathematically 10% of the time (Diaz, 2008).

Some scholars promote a conceptual and problem-solving approach infrequently emphasized in schools serving low-income children (Stipek & Ryan, 1997) that may explicitly support African American students' participation in increasingly sophisticated forms of mathematical communication and argumentation. For example, asking students "How do you know?" as opposed to a more didactic approach of giving information [frequently used with African American students (Haberman, 1991; Jackson & Wilson, 2012; Ladson-Billings, 1997)] may be especially beneficial for African American students.

The Present Study

Data for the current analysis were drawn from an evaluation of the "Technology-enhanced, Research-based, Instruction, Assessment, and professional Development" (TRIAD) model for scaling up successful interventions (Clements et al., 2011; Sarama & Clements, 2013; Sarama et al., 2008). This study assessed an instantiation of the TRIAD model that implemented the *Building Blocks* curriculum (Clements & Sarama, 2008), an empirically validated preschool mathematics program based on developmentally appropriate learning trajectories (see Clements & Sarama, 2008). These learning trajectories were operationalized by a series of "developmental progressions," which identified a given topic or domain. To attain a certain mathematical competence in a given topic or domain, students are taught each successive level in a developmental progression using research-based tasks and instructional strategies.

Building Blocks emphasizes both numeracy and spatial/geometric concepts and procedures through the use of whole and small group activities. In addition to the curriculum, the TRIAD intervention included a software package and extensive teacher support, in the form of training on the software, 13 professional development (PD) sessions before and during the school year, and study-appointed mentors and coaches. During the PD sessions, the training addressed numerous effective teaching practices, including how to increase the cognitive demand of the mathematics taught, how to use formative assessment, and how to encourage cognitive development through the use of learning trajectories (see Clements et al., 2011). Of particular importance to this study, PD administrators also worked with treatment teachers to change their beliefs about the ability of low-income and minority children to learn advanced mathematics.

The goal of the present study was to understand (a) which classroom dimensions were associated with African American and non–African American students' mathematics achievement, and (b) whether classroom dimensions mediated the association between the intervention and students' mathematics achievement at the end of preschool. We focused on the dichotomy of African American and non–African American students because treatment evaluations of the intervention suggest that although Hispanic students made statistically significant gains in mathematics achievement, these gains were not statistically significantly different than those of White students (Clements et al., 2011).

A critical feature of the present study is that we were able to examine the associations of specific dimensions of the classroom (e.g., teacher expectations and developmental appropriateness, teacher confidence and enthusiasm, and high quality instructional practices) within the context of a randomized control trial. Further, we examined whether these dimensions of classroom practices were associated with later measures of student achievement.

Previous research affords the following hypotheses and predictions: We expect teacher expectations and developmental appropriateness will matter more for African American students' achievement than for non–African American students' achievement and help explain (i.e., mediate) the association between the TRIAD treatment and African American students' mathematics achievement (H1). Teacher confidence and enthusiasm will be associated with students' posttest mathematics achievement—for both African American and non–African American students—and

will be a mediator of the treatment intervention on mathematics achievement (H2). However, in the absence of theory, we do not predict any differences in this association between African American and non–African American students. Finally, we predict that support for mathematical discourse will be statistically significantly associated with mathematics achievement for both groups (African American *and* non–African American students) but, in the absence of theory, will not differentially predict between the groups (H3). These hypotheses were tested using a multigroup multilevel mediation model conducted for each of the three mediators (teacher expectations and developmental appropriateness, teacher confidence and enthusiasm, and high quality instructional practices).

Method

Participants and Procedure

Data for the current analysis were drawn from the TRIAD evaluation—a study that assessed the scale-up and student-level impacts of the TRIAD intervention model, of which a key component was the *Building Blocks* mathematics curriculum (see Clements et al., 2011; Clements, Sarama, Wolfe, & Spitler, 2013; Sarama, Clements, Wolfe, & Spitler, 2012). The TRIAD evaluation recruited 42 low-resource schools in two states (New York and Massachusetts) to participate in the scale-up evaluation of *Building Blocks.* Schools were grouped into eight blocks based on state achievement scores, and then randomly assigned within block to one of two conditions: (a) *Building Blocks* curriculum condition (treatment), or (b) business-as-usual (control).[1] In the following school year (2006–2007), student-level data collection began, and 1,375 preschool students attending the 42 study-schools were recruited for study participation at the beginning of the school year.

The majority of students in the study qualified for free or reduced price lunch (84% of those for whom we had available data for), and 55% identified as African American and 22% as Hispanic. The analysis sample only included students who had valid baseline and posttreatment test score data (n = 1,305 students), and were in classrooms that were observed on the observation protocol resulting in a final sample size of 1,238 students in 101 classrooms (approximately 70% of the students in each classroom). Twenty-nine percent of the sample was in the control group (n = 350 students), and 71% of the sample was in the treatment group (n = 888 students). Table 1 shows the descriptive statistics for both student- and classroom-level variables for the overall sample and by group (African American and non–African American students).

Student-Level Measures

Mathematics achievement. Children's mathematical knowledge was assessed at preschool entry (pretest) and at the end of the preschool year (posttest) using the Research-based Elementary Math Assessment (REMA; Clements, Sarama, & Liu, 2008). The REMA was designed to assess the mathematics knowledge of children from ages three to eight, specifically students' number (e.g., object counting, number comparison, numeral recognition) and geometry (e.g., shape identification, measurement, patterning) skills, and was administered in two one-on-one sessions where a research assistant verbally asked participants to respond to each item. After the child incorrectly answered four items in a row, the assessment stopped. The assessments were videotaped and later coded by a team of trained researchers for correctness and strategy use. Using these dichotomous scores, Rasch analysis—where item difficulty is considered in estimating a child's overall score—was employed. Using data from multiple years, the scores were then placed on a vertical scale using data from students in grades preschool, kindergarten, first grade, and second grade where first grade was used as the benchmark. Final scores were scaled to have a mean of 0 and a standard deviation of 1.

The REMA has been validated across three diverse samples of preschool-aged children and produced an overall reliability of .93 (see Clements et al., 2008). Further, it has been shown to have a .86 correlation with another empirically validated measure of early mathematics achievement, the Child Math Assessment: Preschool Battery (Klein, Starkey, & Wakeley, 2000) and a .74 correlation with the Woodcock-Johnson Applied Problems subtest (Clements et al., 2008). In the current sample, the measure had a reliability (alpha) of .92.

Student-level covariates. Information regarding student ethnicity, gender, age, and limited English proficiency status were collected from the schools at the beginning of the preschool year.

Classroom-Level Measures

Classroom observations. Live classroom observations were conducted twice during the year using the Classroom Observation of Early Mathematics Environment and Teaching (COEMET; Clements & Sarama, 2000/2016). Observation ratings from both time points (the first observation was conducted between October and December and the second observation between March and May of the school year) were combined to obtain an average rating of the classroom environment during the school year. Observers spent about half a day in each classroom from before the children arrived until right before lunch and were blind to experimental condition. The COEMET is divided into two sections. The first section asked observers to document the number of computers in the classroom and the start and end time of all activities that took place during the observation using interval coding. The second section of the COEMET included broader measures rated during each mathematics activity and contained 28 indicators measuring constructs such as *supporting children's conceptual understanding, teaching strategies,* and *expectations.* Observers coded each item on a 5-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (5). Ratings for each mathematics activity were averaged to obtain a mean score across the 28 indicators. Interrater reliability for the COEMET, computed via simultaneous classroom visits by pairs of observers (10% of all observations, with pair memberships rotated), was 88% (i.e., 88% of the 28 Likert items were coded the same by both assessors); of the 12% of disagreements, 99% were of the same polarity (i.e., if one was agree, the other was strongly agree). Coefficient alpha (interitem correlations) for the two in-

[1] The study actually employed three conditions, two treatment conditions and one control condition. However, during the preschool year, the two treatment conditions did not differ, and thus were combined into one condition in the current analysis.

Table 1
Descriptives for Student-Level and Classroom-Level Analysis Sample (N = 1,238 Students in 101 Classrooms)

Level of analysis	Overall sample (N = 1,238)				African American (N = 675)				Non–African American (N = 563)			
	Mean	*SD*	Min	Max	Mean	*SD*	Min	Max	Mean	*SD*	Min	Max
Student-level												
Pre-math	−3.23	.82	−7.2	−.4	−3.31	.77	−7.2	−.6	−3.13	.88	−7.2	−.4
Post-math	−1.97	.70	−4.9	.6	−2.10	.68	−4.9	.6	−1.82	.70	−4.8	.2
Age	4.34	.35	3.7	6.7	4.27	.33	3.7	5.8	4.41	.36	3.7	6.7
African American	55%				—				—			
White	18%				—				39%			
Hispanic	22%				—				48%			
Other ethnicity	5%				—				13%			
LEP	17%				3%				33%			
Special education	17%				15%				20%			
Male	49%				50%				49%			

	Overall sample (N = 101)				Treatment (N = 888 students in 69 classrooms)				Control (N = 350 students in 32 classrooms)			
	Mean	*SD*	Min	Max	Mean	*SD*	Min	Max	Mean	*SD*	Min	Max
Classroom-level												
No. computers	2.06	1.32	0.0	6.0	2.45	1.21	0.0	6.0	1.22	1.16	0.0	4.0
No. math activities	6.97	2.85	1.0	15.0	7.75	2.64	2.0	15.0	5.28	2.58	1.0	12.0
Minutes of math	30.90	14.60	8.2	92.5	32.39	15.39	8.2	92.5	27.68	12.33	10.3	60.7
Homework	2.40	1.06	1.0	4.0	2.51	1.02	1.0	4.0	2.19	1.12	1.0	4.0
Classroom size	17.07	3.18	10.0	25.0	17.48	2.98	11.0	25.0	16.22	3.45	10.0	22.0
Master's degree	0.82	0.39	0.0	1.0	0.88	.33	0.0	1.0	0.69	0.47	0.0	1.0
No. years teaching	15.29	8.61	0.0	35.0	15.10	8.35	1.0	32.0	15.69	9.27	0.0	35.0
PD hours	14.90	8.49	0.0	342.0	15.10	8.35	0.0	32.0	14.45	8.93	0.0	35.0
Percent African American	0.54	0.35	0.0	1.0	0.57	0.35	0.0	1.0	0.49	0.34	0.0	1.0
Percent SES	0.38	0.19	0.0	0.9	0.40	0.18	0.1	0.9	.33	0.19	0.0	0.7
Confidence	3.01	0.37	1.8	3.8	3.09	0.30	2.0	3.6	2.84	0.46	1.8	3.8
Expectations	3.15	0.21	1.9	3.8	3.19	0.16	2.4	3.8	3.09	0.27	1.9	3.4
Discourse	3.00	0.38	1.9	4.1	3.07	0.34	2.4	4.1	2.85	0.43	1.9	3.6

Note. Mathematics achievement was scored using Rasch analysis on a vertical scale. PD = professional development; SES = socioeconomic status; LEP = limited English proficient. PD hours is the number of hours of professional development before the intervention. No. computers is the number of computers. No. math activities is number of math activities. No. years teaching is number of years teaching. Difference in the number of computers and number of discrete mathematics activities were statistically significantly different across the treatment and control groups at $p < .001$.

struments ranged from .95 to .97 in previous research (Clements & Sarama, 2008; Clements et al., 2011).

Three dimensions of classroom quality were used in the analysis: *expectations and developmental appropriateness; teacher confidence and enthusiasm; and support for mathematical discourse.* Information on the factor analyses, items, and standardized factor loadings is available in the supplemental materials (Tables S1 through S3). Correlations among factors were moderate in magnitude: r = .65 between teacher confidence and enthusiasm and expectations and responsiveness to developmental needs; r = .48 between teacher confidence and enthusiasm and support for mathematical discourse; r = .66 between expectations and responsiveness to developmental needs and support for mathematical discourse.

Expectations and responsiveness to developmental needs. Eight items were used to describe the extent to which the teacher had high expectations for her students and whether or not the teacher engaged with children in a developmentally appropriate manner. Developmental appropriateness was operationalized by the researchers as the extent to which teachers engaged in practices that matched the developmental abilities and potential of the students. This meant that a teacher who was rated high on developmental appropriateness engaged students in higher-level thinking around mathematics and English language arts instead of engaging students in only play-based activities, with no emphasis on higher-level thinking. Standardized factor loadings ranged from .73 to .88 (α = .94). We created a composite using items weighted by their standardized factor loadings.

Teacher confidence and enthusiasm. Three items were used to describe characteristics that were associated with how confident and enthusiastic the teacher appeared during class. Specifically, items described if the teacher was confident in her teaching, showed interest and value in her students, and displayed enthusiasm for mathematics. Standardized factor loadings ranged from .68 to .84 (α = .82). We created a composite score using items weighted by their standardized factor loadings for the analysis.

Support for mathematical discourse. Eight items characterized the specific mathematics practices and strategies the teacher used during the observed lessons. These items captured the extent to which the teacher elicited higher-order thinking in mathematics through supporting students' explanations and thinking. Standardized factor loadings ranged from .82 to .88 (α = .94). We created a composite using items weighted by their standardized factor loadings.

Classroom-level covariates. Classroom-level covariates came from three sources: (a) the teacher survey, (b) classroom observations, and (c) student demographic information. Study teachers were administered a survey twice during the school year, in the fall and spring of the study year. For teachers who did not respond to the fall survey (two teachers), we used their responses from the spring survey. We included information on whether the teacher held a master's degree, the number of students in the classroom, the number of years of preschool through Grade 12 teaching experience, the number of PD hours prior to the intervention, and whether the teacher assigned mathematics homework measured on a one to four Likert scale with anchor points *almost never* to *a lot*. From the classroom observations, we controlled for the number of computers in the classroom, number of minutes spent on mathematics, and the number of discrete mathematics tasks observed. Number of minutes spent on mathematics was the observed actual time of mathematics instruction from the recorded start and end times on the COEMET. The number of discrete mathematics tasks observed was obtained from the interval coding of the COEMET where observers had to start coding at the beginning of each mathematics activity.

Finally, we aggregated information from the student demographic information to include percent African American students in the classroom as well as percent of students' whose mother's education was college or higher as a proxy for classroom-level socioeconomic status (SES). Because of constraints on researcher resources, only data from approximately 70% of the students in each classroom were collected. As such, the classroom-aggregated measures serve as approximations of the average percent of African American students and average SES of the classroom. The racial/ethnic breakdown of the classroom composition varied considerably. The mean percent of African American students in a classroom was 54% ranging from classrooms with no African American students (0%) to classrooms with only African American students (100%).

Overview of Analytic Models and Method

To examine differences in mediational paths between African American and non–African American students, we conducted a multigroup multilevel meditational analysis (Asparouhov & Muthén, 2012; Retelsdorf, Schwartz, & Asbrock, 2015). With this method, we combined a multilevel structural equation model examining mediation effects (Preacher, Zyphur, & Zhang, 2010) with multigroup analysis in *Mplus 7.2* (Muthén & Muthén, 2013). We specified a 2→2→1 model of mediation in which classroom observation (level 2) mediated the association between treatment (level 2) and student mathematics achievement (level 1). The equations for the direct effect (path c) of the treatment on post mathematics achievement are:

Level-1 (student-level) equation:

$$PostMath_{ij} = \beta_{0j} + \beta_1 PreMath_{ij} + \lambda_2 Covariates_{ij} + r_{ij}$$

Level-2 (classroom-level) equation:

$$\beta_{0j} = \gamma_{00} + \gamma_c Treatment_j + \lambda_2 Covariates_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

where ($PostMath_{ij}$) is the posttreatment measure of mathematics achievement for the *i*th student in school *j*. Posttreatment mathematics achievement is modeled as a function of an overall intercept (γ_{00}, the main effect for the treatment group ($\gamma_c Treatment_j$), a vector of level-2 covariates ($\gamma_2 Covariates_j$), the students' baseline mathematics achievement ($\beta_1 PreMath_{ij}$), a vector of level one covariates ($\gamma_2 Covariates_j$), and a classroom-level (u_{0j}) and student-level (r_{ij}) error term.

The equations for the effect of the mediator on posttreatment mathematics achievement (path b) are the same as above, except that a mediator component is added to the level-2 equation:

Level-1 (student-level) equation:

$$PostMath_{ij} = \beta_{0j} + \beta_1 PreMath_{ij} + \lambda_{2j} Covariates_{ij} + r_{ij}$$

Level-2 (classroom-level) equation:

$$\beta_{0j} = \gamma_{00} + \gamma_{c'} Treatment_j + \gamma_b Mediator_j + \lambda_3 Covariates_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

The equation for the effect of the treatment on the mediator (path a) is:

$$Mediator_j = \beta_0 + \beta_a Treatment_j + r_j$$

where the mediator is a function of an overall intercept (β_0), the main effect of the treatment ($\beta_a Treatment_j$), and a classroom-level error term (r_j).

Mediation is implied if the treatment is significantly related to the mediator (classroom observation), the mediator is significantly associated with the outcome (student achievement), and the mediator accounts for a significant portion of the variance of the association of the treatment on the outcome. Further, we also specified a multiple group model whereby the effect of classroom observation (level 2) on student mathematics achievement (level 1) differed between African American and non–African American students. Because our grouping variable varied within classroom, calculating varying effects for African American and non–African American students cannot be directly specified because the estimated variance/covariance matrix for the observed variables will be group-specific (i.e., classroom specific). However, the effect of the classroom observation on mathematics achievement is correlated between racial/ethnic groups because these students are in the same classroom, but the effect need not be the same for both ethnicities.

To circumvent this within the multilevel structural equation framework, we used a method proposed by Asparouhov and Muthén (2012) where latent variables are introduced in the model to account for the covariance between the group specific classroom effects (please refer to the supplemental materials for our *Mplus* code). Specifically, known latent classes using race (African American and non–African American) were specified in the model such that separate effects for African American and non–African American students are estimated. In the between portion of the model where level 2 effects are specified, two latent variables (our known latent classes) are introduced to represent the between level random effects for $e_{non\text{–}African\ American}$ and $e_{African\ American}$. The math outcome was specified to have a zero residual variance structure as well as loadings equal to one in both groups. The random effects were correlated within classroom to account for nesting. This results in a model where the outcome (posttreatment mathematics achievement) was represented by $e_{non\text{–}African\ American}$ and $e_{African\ American}$. A Wald test was used to examine differences

in the b paths (path between classroom dimension and math achievement) between the two groups. To estimate the indirect effects and report a statistical test for mediation, we used a parametric bootstrap method (Efron & Tibshirani, 1986) in which the parameter point estimates for the indirect effect are generated from random draws of the parameter distributions for the a and b paths of the mediation model. This method has advantages over the Sobel test in that the distribution of the indirect effect is not assumed to be normally distributed (Preacher et al., 2010). To estimate the parametric bootstrap, we used a Web-based tool to generate R code developed by Selig and Preacher (2008) specifying a confidence interval of 95% and 20,000 random draws. The p values for the average indirect effects were then calculated directly in R and reported. In total, three models were run—one for each mediator.

Results

We first present results from the correlations among the variables of interest, which are displayed in Table 2. We then discuss results from each of the mediation models with the three classroom observation variables for non–African American and African American groups. These are displayed as figures (Figures 1 through 3) depicting path diagrams as well as in table form (Tables 3 through 5). We have also included information on the same mediation models comparing White students with Hispanic students ($n = 488$) to further justify why we collapsed these two racial/ethnic groups into the non–African American group. These results are described at the end of each of the mediator sections, and full tables can be found in the supplemental materials (Tables S4 through S6). Wald tests were only performed if the path from the classroom dimension to students' mathematics achievement was statistically significant for at least one of the groups.

Baseline mathematics achievement and posttreatment mathematics achievement were highly correlated, $r = .57$, $p < .001$. Across the whole sample, teacher confidence and enthusiasm was statistically significantly correlated with post mathematics achievement ($r = .18$, $p < .001$), as was expectations and responsiveness to developmental needs ($r = .09$, $p < .01$), but, support for mathematical discourse and posttest mathematics achievement were not significantly correlated, $r = -.03$, $p = .29$.

Number of computers in the classroom (recall that the *Building Blocks* intervention included software) was statistically significantly correlated with teacher confidence and enthusiasm and expectations and responsiveness to developmental needs ($r = .22$, $p < .001$; $r = .14$, $p < .001$, respectively) but not with support for mathematical discourse, $r = -.02$, $p = .48$. Number of mathematics activities was also significantly correlated with teacher confidence and enthusiasm and expectations and responsiveness to developmental needs ($r = .31$, $p < .001$; $r = .11$, $p < .001$, respectively) but significantly negatively correlated with support for mathematical discourse, $r = -.26$, $p < .001$. Whether the teacher assigned mathematics homework was significantly negatively correlated with all three classroom-observation variables ($r = -.08$, $p < .01$; $r = -.10$, $p < .001$; $r = -.28$, $p < .001$; for teacher confidence and enthusiasm, expectations and responsiveness to developmental needs, and support for mathematical discourse, respectively).

Mediator 1: Expectations and Responsiveness to Developmental Needs

Figure 1 displays the path diagram for the first mediator: expectations and responsiveness to developmental needs. The treatment had a positive impact on teacher expectations ($b = 0.10$, $p = .05$). For non–African American students the path from expectations and responsiveness to developmental needs to post mathematics achievement was not statistically significantly different from zero ($b = -0.08$, $p = .57$) and there was no mediation ($b = -0.01$, $p = .33$, 95% CI [−.05, .02]). For African American students, expectations and responsiveness to developmental needs was significantly associated with post mathematics achievement ($b = 0.54$, $p < .001$). However, the parametric bootstrap test for mediation only approached statistical significance ($b = 0.05$, $p = .07$, 95% CI [0.002, 0.12]). These findings suggest that African American students especially benefited from teacher expectations and responsiveness to developmental needs.

A Wald test comparing the point estimates of the path from expectations and responsiveness to developmental needs of the two groups (African American and non–African American students) was statistically significant ($p < .001$). The coefficients, standard errors, p values, and 95% confidence intervals for the full list of variables are displayed in Table 3.

In the model comparing White and Hispanic students (488 students in 88 classrooms), the treatment did not quite have a significant impact on teacher expectations ($b = .09$ $p = .06$). For Hispanic students, the path from expectations and responsiveness to developmental needs to post mathematics achievement was not significantly different from zero ($b = -0.21$, $p = .33$). For White students, expectations and responsiveness to developmental needs was not significantly associated with post mathematics achievement ($b = 0.03$, $p = .93$). The parametric bootstrap test for mediation was not statistically significant for either group. The full table (Table S1) of results is displayed in the supplemental materials.

Mediator 2: Teacher Confidence and Enthusiasm

Figure 2 displays the path diagram for the second mediator: teacher confidence and enthusiasm. Teachers in the treatment group on average were rated 0.25 points higher ($p = .001$) on teacher confidence and enthusiasm than teachers in the control group.

For non–African American students, the path from teacher confidence and enthusiasm to post mathematics achievement was not significantly different from zero ($b = 0.14$, $p = .12$). The parametric bootstrap test indicated no statistically significant mediation for non–African American students ($b = 0.04$, $p = .10$, 95% CI [0.0003, 0.08]). For African American students, teacher confidence and enthusiasm was statistically significantly associated with posttreatment mathematics achievement ($b = 0.25$, $p = .001$) such that a one-unit increase in observed teacher confidence and enthusiasm yielded a predicted 0.25 unit increase in their posttreatment mathematics scores. The parametric bootstrap test indicated partial mediation for African American students ($b = 0.06$, $p = .03$, 95% CI [0.01, 0.13]). This implies that part of the effects of the treatment were through increasing teacher's confidence and enthusiasm for teaching.

Table 2

Correlations Among Classroom Observations, Teacher Characteristics, and Students' Mathematics Achievement in 101 Classrooms (N = 1,238)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. T1 math	1.00																			
2. T2 math	**.57**	1.00																		
3. Male	−.06	−.06	1.00																	
4. African American	**−.11**	**−.20**	.02	1.00																
5. Hispanic	**−.10**	.03	−.03	**−.58**	1.00															
6. LEP	**−.14**	.01	−.01	**−.40**	**.50**	1.00														
7. Special education	−.03	−.04	.01	−.07	.07	−.01	1.00													
8. Age	**.29**	**.30**	.05	**−.20**	**.25**	**.24**	**.08**	1.00												
9. L2 confidence	**.09**	**.19**	−.01	**.08**	**−.12**	**−.18**	.03	**−.09**	1.00											
10. L2 expectations	**.08**	**.09**	−.04	**.12**	**−.28**	**−.29**	.01	**−.19**	**.60**	1.00										
11. L2 discourse	.02	−.03	−.03	**.20**	**−.36**	**−.36**	−.05	**−.30**	**.43**	**.62**	1.00									
12. L2 number of computers	.01	**.18**	.02	**−.10**	**.08**	**.09**	.02	−.03	**.22**	**.14**	−.02	1.00								
13. L2 number of math activities	.05	**.27**	−.01	**−.08**	**.13**	**.11**	.05	**.17**	**.31**	**.11**	**−.26**	**.44**	1.00							
14. L2 minutes of math	.06	**.11**	.02	.03	**−.10**	**−.08**	−.02	.02	**.24**	**.12**	−.02	**.14**	**.38**	1.00						
15. L2 assign math homework	−.01	**.10**	−.05	.03	**.11**	**.12**	.05	**.10**	**−.08**	**−.10**	**−.28**	**.19**	**.35**	.01	1.00					
16. L2 classroom size	.02	**.10**	−.01	**−.13**	**.22**	**.24**	.05	**.22**	.06	**−.10**	**−.19**	**.10**	**.31**	.02	**.21**	1.00				
17. L2 Master's degree	.00	−.02	−.00	**.17**	**−.33**	**−.24**	−.04	**−.23**	**.18**	**.29**	**.44**	.02	**−.19**	**.13**	−.05	**−.27**	1.00			
18. L2 No. years teaching	−.06	−.04	.02	**.11**	−.03	**−.08**	−.02	**−.11**	**−.09**	**−.09**	−.05	.03	−.07	−.05	−.04	**−.20**	**.08**	1.00		
19. L2 PD hours	−.06	−.01	.03	**.13**	**−.08**	**−.08**	−.01	**−.11**	**−.08**	**−.08**	−.03	.07	−.03	−.04	.00	**−.17**	**.20**	**.96**	1.00	
20. L2 African American	**−.08**	**−.13**	.04	**.69**	**−.49**	**−.43**	−.06	**−.22**	**.12**	**.18**	**.29**	**−.14**	**−.12**	.05	.04	**−.18**	**.25**	**.17**	**.19**	1.00
21. Percent SES	**.14**	**.13**	−.03	−.01	**−.09**	**−.18**	.04	**−.11**	**.23**	**.14**	.05	**.25**	**.08**	−.01	**.11**	.00	**.17**	.00	.07	−.01

Note. LEP = limited English proficient; PD = professional development; SES = socioeconomic status. Bolded correlations are statistically significant at $p < .01$. L2 is classroom level.

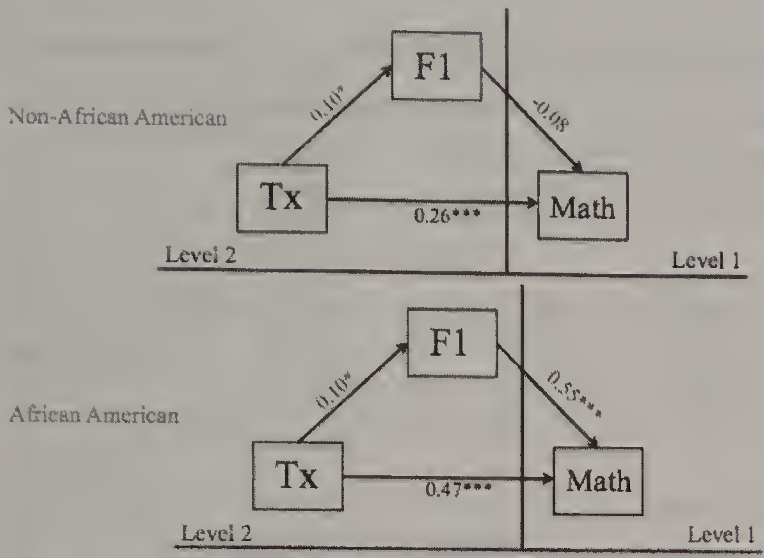


Figure 1. Multigroup multilevel mediation for expectations and responsiveness to developmental needs. Unstandardized regression coefficients at the between-level are shown. Wald test was statistically significant. Results from the parametric bootstrap test for mediation for the African American group suggest partial mediation approaching statistical significance. Tx = treatment; F1 = expectations and responsiveness to developmental needs. $^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

Although we found evidence of partial mediation for African American students but not for non–African American students, a Wald test comparing the point estimates of the path from teacher confidence and enthusiasm to post mathematics achievement was not statistically significant (p = .28). It should be noted that the confidence interval for the point estimate for the non–African American group includes the point estimate from the African American group suggesting that the effects for the two groups might not actually differ. As such, we suggest that limited evi-

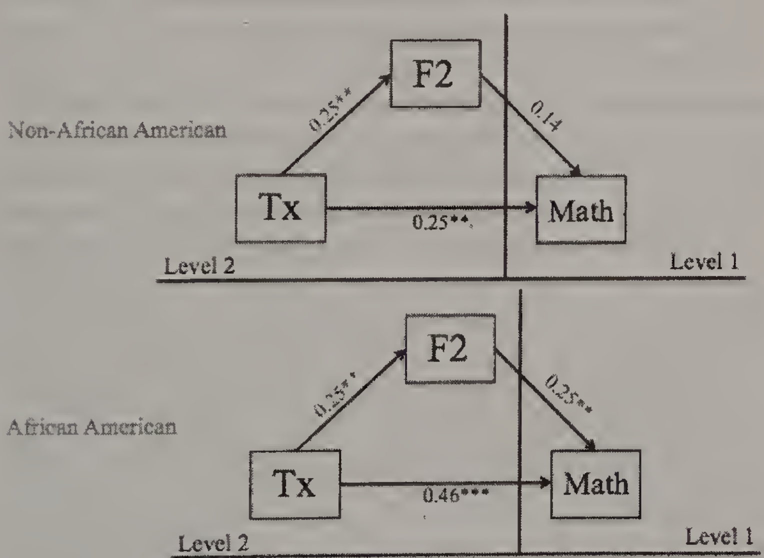


Figure 2. Multigroup multilevel mediation for teacher confidence and enthusiasm. Unstandardized regression coefficients at the between-level are shown. Wald test was not statistically significant. Results from the parametric bootstrap test for mediation suggest partial mediation for African American students but no mediation for non–African American students. Tx = treatment; F2 = teacher confidence and enthusiasm. $^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

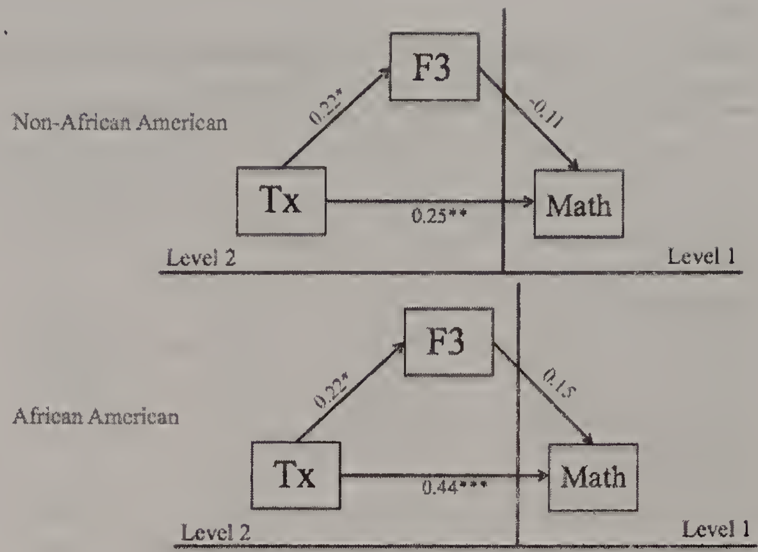


Figure 3. Multigroup multilevel mediation for support for mathematical discourse. Unstandardized regression coefficients at the between-level are shown. No significant mediation. Tx = treatment; F3 = support for mathematical discourse. $^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

dence supports the conclusion that teacher confidence and enthusiasm differentially influenced African American students' mathematics achievement. The coefficients, standard errors, p values, and 95% confidence intervals for the full list of variables are displayed in Table 4.

In the model comparing White and Hispanic students (488 students in 88 classrooms), the treatment had a positive impact on teacher confidence (b = .20, p = .02). For Hispanic students the path from teacher confidence to post mathematics achievement was not statistically significantly different from zero (b = 0.06, p = .68). For White students, teacher confidence was also not statistically significantly associated with post mathematics achievement (b = 0.19, p = .12). The parametric bootstrap test was not statistically significant for either group. The full table (Table S2) of results is displayed in the supplemental materials.

Mediator 3: Support for Mathematical Discourse

Figure 3 displays the path diagram for the third mediator: support for mathematical discourse. Teachers in the treatment group on average were rated 0.22 points higher (p = .01) on support for mathematical discourse than teachers in the control group.

The path from support for mathematical discourse to posttreatment mathematics achievement was not statistically significant for both groups (b = −0.11, p = .24; b = 0.16, p = .13, for non–African American and African American students, respectively). As such, the parametric bootstrap test was not statistically significant for either group (b = −0.02, p = .46, 95% CI [−0.08, 0.02]; b = 0.03, p = .22, 95% CI [−0.01, 0.10], for non–African American and African American students, respectively). The coefficients, standard errors, p values, and 95% confidence intervals for the full list of variables are displayed in Table 5.

In the model comparing White and Hispanic students (488 students in 88 classrooms), the treatment had a positive impact on support for mathematical discourse (b = .21, p = .02). For Hispanic students, the path from mathematical discourse to post

Table 3
Results From the Multigroup Multilevel Mediation Model With Expectations and Responsiveness to Developmental Needs as the Mediator (N = 1,238 Students in 101 Classrooms)

Mediator	African American students			Non–African American students		
	β (*SE*)	95% CI	*p*	β (*SE*)	95% CI	*p*
Indirect effect (Treatment → Post-mathematics)						
Within level						
Pre-mathematics	.44 (.03)	[.40–.49]	<.001	.44 (.03)	[.40–.49]	<.001
Male	−.05 (.03)	[−.10–.00]	.12	−.05 (.03)	[−.10–.00]	.12
Age	.26 (.05)	[.18–.35]	<.001	.26 (.05)	[.18–.35]	<.001
LEP	.07 (.05)	[−.01–.16]	.14	.07 (.05)	[−.01–.16]	.14
Special education	−.07 (.04)	[−.14–−.01]	.07	−.07 (.04)	[−.14–−.01]	.07
Hispanic	—	—	—	−.11 (.05)	[−.20–−.03]	.03
Between level						
Treatment	.47 (.07)	[.37–.58]	<.001	.26 (.07)	[.14–.37]	<.001
Treatment → Expectations and responsiveness to developmental needs						
Between level						
Treatment	.10 (.05)	[.02–.18]	.05	.10 (.05)	[.02–.18]	.05
Expectations and responsiveness to developmental needs → Post-mathematics						
Between Level						
Expectations	.55 (.13)	[.33–.77]	<.001	−.08 (.15)	[−.33–.16]	.57
Number of computers	.00 (.02)	[−.03–.03]	.98	.00 (.02)	[−.03–.03]	.98
Number of math activities	.03 (.01)	[.01−.05]	.003	.03 (.01)	[.01–.05]	.003
Minutes of math	−.001 (.002)	[−.003–.002]	.69	−.001 (.002)	[−.003–.002]	.69
Assign math homework	.02 (.02)	[−.03–.02]	.54	.02 (.02)	[−.03–.02]	.54
Classroom size	−.01 (.01)	[−.02–.01]	.31	−.01 (.01)	[−.02–.01]	.31
Master's degree	−.03 (.07)	[−.13–.08]	.7	−.03 (.07)	[−.13–.08]	.7
Number of years teaching	−.01 (.01)	[−.02–.01]	.41	−.01 (.01)	[−.02–.01]	.41
PD hours	.01 (.01)	[−.01–.02]	.35	.01 (.01)	[−.01–.02]	.35
Percent African American	.06 (.09)	[−.11–.15]	.50	.06 (.09)	[−.11–.15]	.50
Percent SES	−.01 (.12)	[−.20–.19]	.64	−.01 (.12)	[−.20–.19]	.64

Note. Unstandardized coefficients are reported. PD = professional development; SES = socioeconomic status; *SE* = standard error; CI = confidence interval; LEP = limited English proficient. Expectations is expectations and responsiveness to developmental needs. PD hours is number of PD hours before the intervention.

mathematics achievement was not statistically different from zero ($b = -0.22$, $p = .11$). For White students, support for mathematical discourse was also not significantly associated with post mathematics achievement ($b = -0.02$, $p = .86$). The parametric bootstrap test for mediation was not statistically significant for both groups. The full table (Table S3) of results is displayed in the supplemental materials.

Discussion

This study evaluated the differential effects of three instructional practices on African American students' achievement in the context of a randomized control trial of preschool classrooms. Specifically, we found that, on average, African American students benefited more from certain instructional practices (teacher expectations and responsiveness to developmental needs) than non–African American students and that teacher expectations and responsiveness to developmental needs only approached significance for the partially mediated effect of the intervention on African American students' mathematics achievement. No support was found for the effect of support for mathematical discourse on student achievement for either group.

In this study, we compared African American students with non–African American students such that White *and* Hispanic students were included in the former category. This decision was further justified when we replicated our analyses and restricted the sample to just those two groups and found no statistically significant differences between the groups. In recent years, Hispanic and African American children have shown different achievement trajectories relative to White students (National Center for Education Statistics, 2013; Reardon, Robinson-Cimpian, & Weathers, 2015; for an exception see Rumberger & Palardy, 2005). Whereas White-African American achievement gaps increase during the first six years of schooling in both math and reading, White-Hispanic gaps decrease during this period (Fryer & Levitt, 2004, 2006; Hemphill & Vanneman, 2011; Reardon & Galindo, 2006; Reardon & Robinson, 2008). Our findings are consistent with Bodovski and Farkas (2007) and Bottia and colleagues (2014) who reported modest reductions in the achievement gap between African American and White students in kindergarten as a result of certain instructional practices, but no reductions in the achievement gap between Whites and Hispanics. For example, instruction that focuses on developmentally appropriate but higher-level thinking has been shown to have a positive impact on African American but not Hispanic children's mathematics performance (Bodovski & Farkas, 2007). It is unclear from the empirical and theoretical literature why African American and Hispanic children responded differentially to the classroom environment. Theory

Table 4

Results From the Multigroup Multilevel Mediation Model With Teacher Confidence and Enthusiasm as the Mediator (N = 1,238 Students in 101 Classrooms)

Mediator	African American students			Non–African American students		
	β (*SE*)	95% CI	*p*	β (*SE*)	95% CI	*p*
Indirect effect (Treatment → Post-mathematics)						
Within level						
Pre-mathematics	.44 (.02)	[.41–.48]	<.001	.44 (.02)	[.41–.48]	<.001
Male	−.05 (.03)	[−.10–.01]	.07	−.05 (.03)	[−.10–.01]	.07
Age	.27 (.05)	[.18–.35]	<.001	.27 (.05)	[.18–.35]	<.001
LEP	.08 (.05)	[−.002–.17]	.11	.08 (.05)	[−.002–.17]	.11
Special education	−.08 (.04)	[−.14–−.01]	.05	−.08 (.04)	[−.14–−.01]	.05
Hispanic	—	—	—	−.10 (.05)	[−.19–−.02]	.05
Between level						
Treatment	.46 (.07)	[.35–.57]	<.001	.25 (.07)	[.13–.37]	.001
Treatment → Teacher confidence and enthusiasm						
Between level						
Treatment	.25 (.08)	[.13–.37]	.001	.25 (.08)	[.13–.37]	.001
Teacher confidence and enthusiasm → Post-mathematics						
Between level						
Confidence	.25 (.08)	[.13–.38]	.001	.14 (.07)	[−.01–.29]	.03
Number of computers	−.003 (.02)	[−.03–.03]	.89	−.003 (.02)	[−.03–.03]	.89
Number of math activities	.02 (.01)	[.002−.04]	.07	.02 (.01)	[.002− .04]	.07
Minutes of math	−.00 (.002)	[−.003–.002]	.79	−.00 (.002)	[−.003–.002]	.79
Assign math homework	.03 (.02)	[−.01–.07]	.23	.03 (.02)	[−.01–.07]	.23
Classroom size	−.01 (.01)	[−.02–.003]	.23	−.01 (.01)	[−.02–.003]	.23
Master's degree	−.06 (.07)	[−.17–.05]	.39	−.06 (.07)	[−.17–.05]	.39
Number of years teaching	−.01 (.01)	[−.02–.01]	.30	−.01 (.01)	[−.02–.01]	.3
PD hours	.01 (.01)	[−.003–.02]	.21	.01 (.01)	[−.003–.02]	.21
Percent African American	.02 (.08)	[−.11–.15]	.77	.02 (.08)	[−.11–.15]	.77
Percent SES	−.01 (.12)	[−.20–.19]	.96	−.01 (.12)	[−.20–.19]	.96

Note. Unstandardized coefficients are reported. PD = professional development; SES = socioeconomic status; *SE* = standard error; CI = Confidence Interval; LEP = limited English proficient. Confidence is teacher confidence and enthusiasm. PD hours is number of PD hours before the intervention.

from García Coll and colleagues (1996) suggests that children of different racial/ethnic groups hold differing cultural values, experiences of prejudice, and family values, which may all be factors in children's perceptions of and participation in the learning environment. This issue is still not well understood and future research is needed to better understand how racial/ethnic background influences students' experiences of the learning environment. It should be noted that in the present study, we only examined three teacher practices and it is quite possible that an examination of other classroom dimensions would yield instances where Hispanic students respond differentially to the classroom environment than White students.

A contribution of this study is that we address a limitation of previous studies investigating racial/ethnic differences in responses to the classroom (e.g., Bodovski & Farkas, 2007; Bottia et al., 2014; Sonnenschein & Galindo, 2015). Specifically, we were able to directly investigate within-classroom variation in responsiveness to instructional practices rather than relying on a dataset that was collected at the student-level. A student-level dataset is problematic because observations or teacher reports of the *same* instructional practices were not collected and it could not be investigated whether students, within the same classroom, differentially benefit from certain instructional practices. It may suggest that other confounding factors such as access to quality classrooms or the actual classroom environment children experience may influence the interpretation of previous studies. Additionally, instructional practices in previous studies (e.g., Bodovski & Farkas, 2007; Bottia et al., 2014; Sonnenschein & Galindo, 2015) were measured using teacher reports, which may have limited reliability.

Teacher Expectations and Developmental Appropriateness

Findings from our study extend the research on teacher expectations and student achievement in the following ways: (a) observed teacher expectations can be modified through professional development implemented as part of a mathematics intervention, (b) teacher expectations differentially influence mathematics achievement for African American students, and (c) teacher expectations significantly influence students' mathematics achievement in young students (mean age 4.34 years) supporting earlier work suggesting the importance of early expectations (Alvidrez & Weinstein, 1999). In support of theory on teacher expectancy beliefs, the effects we found in our sample of preschool students may have been strongest during that year because teachers had little prior contact with children, the teacher was integral in conveying the mathematics knowledge to students, and the mathematics content was novel (West & Anderson, 1976). It may be that African American students respond positively to these kinds of teacher practices and behaviors because these practices and beliefs have been found to be absent from environments (school, home, or otherwise) where large numbers of African Americans participate (Oakes, 1990). Theoretical work on African American students'

Table 5
Results From the Multigroup Multilevel Mediation Model With Support for Mathematical Discourse as the Mediator (N = 1,238 Students in 101 Classrooms)

Mediator	African American students			Non–African American students		
	β (*SE*)	95% CI	*p*	β (*SE*)	95% CI	*p*
Indirect effect (Treatment → Post-mathematics)						
Within level						
Pre-mathematics	.45 (.03)	[.40−.51]	<.001	.45 (.03)	[.40−.51]	<.001
Male	−.05 (.03)	[−.10−.003]	.13	−.05 (.03)	[−.10−.003]	.13
Age	.26 (.05)	[.17−.34]	<.001	.26 (.05)	[.17−.34]	<.001
LEP	.07 (.05)	[−.02−.15]	.2	.07 (.05)	[−.02−.15]	.20
Special education	−.07 (.04)	[−.14−−.01]	.07	−.07 (.04)	[−.14−−.01]	.07
Hispanic	—	—	—	−.12 (.05)	[−.21−−.04]	.02
Between level						
Treatment	.44 (.08)	[.31−.57]	<.001	.25 (.08)	[.12−.38]	.001
Treatment → Support for mathematical discourse						
Between level						
Treatment	.22 (.09)	[.08–.36]	.01	.22 (.09)	[.08–.36]	.01
Support for mathematical discourse → Post-mathematics						
Between level						
Discourse	.15 (.10)	[−.01–.33]	.13	−.11 (.10)	[−.27–.05]	.24
Number of computers	.01 (.02)	[−.03–.04]	.80	.01 (.02)	[−.03–.04]	.80
Number of math activities	.03 (.01)	[.02–.05]	.001	.03 (.01)	[.02–.05]	.001
Minutes of math	.00 (.002)	[−.003–.002]	.83	.00 (.002)	[−.003–.002]	.83
Assign math homework	.01 (.03)	[−.03–.05]	.70	.01 (.03)	[−.03–.05]	.70
Classroom size	−.01 (.01)	[−.02–.004]	.27	−.01 (.01)	[−.02–.004]	.27
Master's degree	−.02 (.07)	[−.14–.10]	.80	−.02 (.07)	[−.14–.10]	.80
Number of years teaching	−.01 (.01)	[−.02–.01]	.44	−.01 (.01)	[−.02–.01]	.44
PD hours	.01 (.01)	[−.01–.02]	.37	.01 (.01)	[−.01–.02]	.37
Percent African American	.05 (.09)	[−.10–.19]	.61	.05 (.09)	[−.10–.19]	.61
Percent SES	.07 (.14)	[−.15–.29]	.60	.07 (.14)	[−.15–.29]	.60

Note. Unstandardized coefficients are reported. PD = professional development; SES = socioeconomic status; *SE* = standard error; CI = Confidence Interval; LEP = limited English proficient. Discourse is support for mathematical discourse. PD hours is number of PD hours before the intervention.

mathematics learning also suggests the expectations teachers have for these students are important (Ladson-Billings, 1997). It may also be that African American students are typically not exposed to environments where individuals have high expectations of them (Ferguson, 2000; Lopez, 2002). However, most of the research suggesting teacher expectations are especially important for African American students is usually qualitative in nature and does not compare the effects of teacher expectations on achievement across different racial/ethnic groups. The present study provided empirical support for this assertion but cannot answer why. More research using student and teacher interviews could shed light on this question.

The expectations teachers have of their students may be an indicator of overall classroom quality. For example, Rubie-Davies (2007) found statistically significant differences in observed practices of teachers who were low, average, and high in their expectations of students. Teachers in the high expectations group were observed to provide more feedback to their students, engage in more higher-order questioning, and were rated higher in managing student behavior. In the present study, we found evidence of this in the large and statistically significant correlations among the three classroom dimensions we measured. These findings suggest that efforts to improve teachers' instructional practices could also include professional development opportunities aimed at changing teacher's expectations of their students and can even be implemented at the same time as interventions focused on improving mathematics achievement. This integrated view of mathematics whereby improvement of elements of mathematics instruction *and* the classroom climate is supported in the NCTM standards (NCTM, 2000). Additionally, we note that it is important to consider multiple dimensions of the classroom-learning environment as only looking at teacher expectancy effects could mask our understanding of other important dimensions.

Teacher Confidence and Enthusiasm

The intervention significantly increased the observed teacher confidence and enthusiasm of teachers in the treatment condition. This is not surprising given that the intervention was mathematics-specific and provided treatment teachers with ample support to teach mathematics through the *Building Blocks* curriculum. Additionally, we found some support that teacher confidence and enthusiasm differentially partially mediated the effect of the intervention on African American students' mathematics achievement but not non–African American students' achievement. Even though we found a statistically significant association between observer's ratings of teacher confidence and enthusiasm and African American students' mathematics achievement, a Wald test comparing the coefficients between the two groups was not statistically significant, suggesting that there may not actually be a statistically significant association between teacher confidence and enthusiasm and African American students' mathematics achievement. We caution readers from over interpreting these findings and instead suggest that future research replicate this finding. Previous

research does, however, suggest that teachers' mathematical knowledge for teaching is associated with gains in student achievement over time (Hill, Rowan, & Ball, 2005) and that mathematical knowledge for teaching could be related to how confidently and enthusiastically a teacher teaches.

Support for Mathematical Discourse

Although we found that the intervention significantly increased teacher's observed support for mathematical discourse, we did not find that this dimension was significantly associated with students' posttreatment mathematics achievement. This was surprising and stands in contrast to theory suggesting the importance of support for mathematical discourse and students' mathematics achievement (e.g., Chi, 2000). In the present study, it may be that support for the mathematical discourse dimension did not have an additional effect on students' achievement above and beyond the elements that were already present in the intervention, such as the use and quality of the *Building Blocks* software. Additionally, the age of the students in our sample may be too young to benefit from high quality support for mathematical discourse. Because observations were conducted of the *teacher's* support for mathematical discourse, it is unclear if the preschool students in our sample actually took up this practice and benefited from it.

Limitations and Future Directions

We note several limitations of this work. Unfortunately, we did not have information about the teacher's ethnicity and therefore could not investigate effects based on teacher and student ethnicity match. Conflicting evidence for the importance of teacher–child ethnicity match exists on outcomes such as achievement and school behavioral adjustment. For example, Ewing and Taylor (2009) did not find that teacher–child ethnicity match moderated the association between relationship quality and behavioral adjustment. However, Saft and Pianta (2001) found that teacher–child ethnicity match was related to teacher's expectations of the child. It was not possible for us to examine teacher and student ethnicity match with our data, but it should be a direction for future research.

The design of the present study prevents us from teasing apart what particular aspects of the intervention led to the differential improvements in student performance. It remains unclear whether this might be attributable to the additional support teachers received during the professional development sessions or the novel attention the research team paid to the teachers. A useful direction for future research would be to test the relative effectiveness of each of the intervention components separately in the form of a treatment "dismantling approach" in which only certain components of mathematics interventions already known to be effective are retained. Further, it is important to note that a unique advantage of our study is that we were able to show that teachers' instructional behaviors, such as holding high expectations for all students, can be changed within the context of a mathematics intervention. Future studies should also consider including a second type of intervention, such as a literacy or socioemotional curriculum, to compare the effects of teachers' general participation in the intervention to the *Building Blocks* intervention.

In considering why many of the proposed classroom mediators were nonsignificant, it could be that other unmeasured characteristics of the classroom were stronger mediators of the intervention for African American students. Another speculation is that whereas classroom environments help support the learning and achievement of students, it is really students' participation in those environments that influences their subsequent achievement (e.g., Ing et al., 2014; Ruzek et al., 2016; Skinner & Belmont, 1993). Ing and colleagues (2014), for example, coded video observations of elementary school mathematics classrooms for the quality of students' engagement the classroom. The authors coded for the level of detail students gave in their explanations as well as the quality of students' engagement and found that the quality of students' engagement related to gains in standardized mathematics achievement. Though these analyses are time and researcher intensive, future research should consider incorporating measures of student engagement—whether through observations or self-report—in understanding the relation between instructional characteristics and student achievement.

We examined the importance of classroom learning environments for students' development of mathematics skills; however, we were unable to examine whether sustained high-quality instruction, and in the case of the present study, sustained teacher expectations could aid in closing the achievement gap between African American students and other groups. A possible fruitful area of future research could be to examine the same associations we investigated in this paper but include more information on the learning environments students are subsequently exposed to. Looking at sustained high-quality instruction could help us understand whether these differential effects of the classroom environment remain even in cases where African American students have had teachers with high levels of confidence and enthusiasm or high expectations for them. Earlier work documenting the middle school and high school transition (Hirsch & Rapkin, 1987; Midgley, Feldlaufer, & Eccles, 1989; Wigfield, Eccles, Mac Iver, Reuman, & Midgley, 1991) has found that continuity in students' experiences mattered for their outcomes. For example, Midgley and colleagues (1989) looked at students' transition from elementary school to junior high and found that students who had rated their teacher as low on support in elementary school but went into classrooms where they rated their teacher as high in support experienced higher intrinsic motivation than students who had initially highly supportive teachers and transitioned into classrooms where they were low in support. This would also help us understand possible reasons for the fadeout effect in educational interventions. Will students who are continually in high-quality classrooms sustain gains that are not found in interventions that only last one year? Rubie-Davies and colleagues (2014) investigated the cumulative effects of teachers from kindergarten to fourth grade on students' fourth grade academic achievement and found that the higher the teacher's expectations were relative to the student's actual ability, the higher the student's achievement in fourth grade. Conversely, the lower the teachers' expectations were over time, the lower the student achieved at the end of fourth grade. In understanding students' continuity of classroom experiences, information on the cumulative effects of classroom environments could be better explored.

Conclusion

Taken together, our findings suggest that specific dimensions of the classroom environment differentially influenced African American students' mathematics learning. Specifically, teacher expectations and developmental appropriateness increased the achievement of African

American students. Additionally, we found that a carefully designed and well-implemented preschool mathematics intervention changed multiple dimensions of the classroom environment. If classroom interventions are likely to help close important race-based achievement gaps, then targeting classroom processes, such as teacher expectations, might prove fruitful in making progress toward this goal. Findings from this study can help researchers and practitioners understand the ways in which the organization of classroom learning environments might be structured more suitably to leverage and support the positive mathematics learning experiences of students who stand to benefit from it the most.

References

Abbott-Shim, M., Lambert, R., & McCarty, F. (2000). Structural model of Head Start classroom quality. *Early Childhood Research Quarterly, 15,* 115–134. http://dx.doi.org/10.1016/S0885-2006(99)00037-X

Aikens, N. L., & Barbarin, O. (2008). Socioeconomic differences in reading trajectories: The contribution of family, neighborhood, and school contexts. *Journal of Educational Psychology, 100,* 235–251. http://dx.doi.org/10.1037/0022-0663.100.2.235

Alexander, K. L., Entwisle, D. R., & Horsey, C. S. (1997). From first grade forward: Early foundations of high school dropout. *Sociology of Education, 70,* 87–107. http://dx.doi.org/10.2307/2673158

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91,* 731–746. http://dx.doi.org/10.1037/0022-0663.91.4.731

Askew, M., Brown, M., Rhodes, V., Wiliam, D., & Johnson, D. (1997). *Effective teachers of numeracy: Report of a study carried out for the Teacher Training Agency.* London, UK: King's College, University of London.

Asparouhov, T., & Muthén, B. O. (2012). *Multiple group multilevel analysis* (Mplus Web Notes, No. 16). Retrieved from www.statmodel.com/examples/webnotes/webnote16.pdf

Battey, D. (2013). "Good" mathematics teaching for students of color and those in poverty: The importance of relational interactions within instruction. *Educational Studies in Mathematics, 82,* 125–144. http://dx.doi.org/10.1007/s10649-012-9412-z

Berry, R. Q., III. (2003). Mathematics standards, cultural styles, and learning preferences: The plight and the promise of African American students. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 76,* 244–249. http://dx.doi.org/10.1080/00098650309602013

Bishop-Josef, S. J., & Zigler, E. (2011). The cognitive/academic emphasis versus whole child approach: The 50-year debate. In E. Zigler, W. S. Gilliam, & W. S. Barnett (Eds.), *The pre-K debates. Current controversies and issues* (pp. 83–88). Baltimore, MD: Brookes.

Boaler, J. (1998). Open and closed mathematics: Student experiences and understandings. *Journal for Research in Mathematics Education, 29,* 41–62. http://dx.doi.org/10.2307/749717

Bodovski, K., & Farkas, G. (2007). Do instructional practices contribute to inequality in achievement? The case of mathematics instruction in kindergarten. *Journal of Early Childhood Research, 5,* 301–322. http://dx.doi.org/10.1177/1476718X07080476

Bottia, M. C., Moller, S., Mickelson, R. A., & Stearns, E. (2014). Foundations of mathematics achievement. *The Elementary School Journal, 115,* 124–150. http://dx.doi.org/10.1086/676950

Brooks-Gunn, J., & Duncan, G. J. (1997). The effects of poverty on children. *Future Child, 7,* 55–71.

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist, 41,* 1069–1077. http://dx.doi.org/10.1037/0003-066X.41.10.1069

Brophy, J. E., & Good, T. L. (1970). Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. *Journal of Educational Psychology, 61,* 365–374. http://dx.doi.org/10.1037/h0029908

Bryant, D. M., Burchinal, M., Lau, L. B., & Sparling, J. J. (1994). Family and classroom correlates of Head Start children's developmental outcomes. *Early Childhood Research Quarterly, 9*(3–4), 289–309. http://dx.doi.org/10.1016/0885-2006(94)90011-6

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (pp. 161–238). Hillsdale, NJ: Erlbaum.

Clarke, J., Frazer, E., DiMartino, J., Fisher, P., & Smith, P. (2003). Making learning personal: Educational practices that work. In J. DiMartino, J. Clarke, & D. Wolk (Eds.) *Personalized learning: Preparing high school students to create their futures* (pp. 173–194). Lanham, MD: Scarecrow Press.

Clements, D. H., Baroody, A. J., & Sarama, J. (2013). Background research on early mathematics. *National Governor's Association, Center Project on Early Mathematics.*

Clements, D. H., & Sarama, J. (2016). *COEMET: The Classroom Observation of Early Mathematics Environment and Teaching instrument.* Denver, CO: University of Denver. (Original work published 2000)

Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculumml: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education, 38,* 136–163.

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal, 45,* 443–494. http://dx.doi.org/10.3102/0002831207312908

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-Based Early Maths Assessment. *Educational Psychology, 28,* 457–482. http://dx.doi.org/10.1080/01443410701777272

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education, 42,* 127–166.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies Persistence of effects in the third year. *American Educational Research Journal, 50,* 812–850. http://dx.doi.org/10.3102/0002831212469270

Darling-Hammond, L. (1997). *The right to learn: A blueprint for creating schools that work.* San Francisco, CA: Jossey-Bass.

Diaz, R. M. (2008). *The role of language in early childhood mathematics* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database (UMI No. 3319002).

Duncan, G. J., Brooks-Gunn, J., & Klebanov, P. K. (1994). Economic deprivation and early childhood development. *Child Development, 65,* 296–318.

Duncan, G., & Magnuson, K. (2005). Can family socioeconomic resources account for racial and ethnic test score gaps? *The Future of Children, 15,* 35–54. http://dx.doi.org/10.1353/foc.2005.0004

Dunn, L., & Kontos, S. (1997). What have we learned about developmentally appropriate practice? Research in review. *Young Children, 52,* 4–13.

Dusek, J. B., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology, 75,* 327–346. http://dx.doi.org/10.1037/0022-0663.75.3.327

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1,* 54–75. http://dx.doi.org/10.1214/ss/1177013815

Evertson, C. M., Emmer, E. T., & Brophy, J. E. (1980). Predictors of effective teaching in junior high mathematics classrooms. *Journal for Research in Mathematics Education, 11,* 167–178. http://dx.doi.org/10.2307/748938

Ewing, A. R., & Taylor, A. R. (2009). The role of child gender and ethnicity in teacher–child relationship quality and children's behavioral adjustment in preschool. *Early Childhood Research Quarterly, 24,* 92–105. http://dx.doi.org/10.1016/j.ecresq.2008.09.002

Ferguson, A. A. (2000). *Bad boys: Public schools in the making of black masculinity.* Ann Arbor, MI: University of Michigan Press. http://dx.doi.org/10.3998/mpub.16801

Ferguson, R. F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation, 28,* 465–498.

Franke, M. L., Fennema, E., & Carpenter, T. (1997). Teachers creating change: Examining evolving beliefs and classroom practice. *Mathematics Teachers in Transition,* 255–282.

Fryer, R. G., Jr., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *The Review of Economics and Statistics, 86,* 447–464. http://dx.doi.org/10.1162/003465304323031049

Fryer, R. G., & Levitt, S. D. (2006). The black-white test score gap through third grade. *American Law and Economics Review, 8,* 249–281. http://dx.doi.org/10.1093/aler/ahl003

García Coll, C., Lamberty, G., Jenkins, R., McAdoo, H. P., Crnic, K., Wasik, B. H., & Vázquez García, H. (1996). An integrative model for the study of developmental competencies in minority children. *Child Development, 67,* 1891–1914. http://dx.doi.org/10.2307/1131600

Gill, S., & Reynolds, A. J. (1999). Educational expectations and school achievement of urban African American children. *Journal of School Psychology, 37,* 403–424. http://dx.doi.org/10.1016/S0022-4405(99)00027-8

Ginsburg, H. P., Inoue, N., & Seo, K. (1999). Young children doing mathematics: Observations of everyday mathematics. In J. Copley (Ed.), *Mathematics in the early years* (pp. 88–100). Washington, DC: National Association for the Education of Young Children.

Ginsburg, H., Lee, J., & Boyd, J. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report,* XX11, 1. Retrieved from http://www.scrd.org

Haberman, M. (1991). The pedagogy of poverty versus good teaching. *Phi Delta Kappan, 73,* 290–294.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale* (rev. ed.). New York, NY: Teachers College Press.

Hemphill, F. C., & Vanneman, A. (2011). *Achievement gaps: How Hispanic and white students in public schools perform in mathematics and reading on the National Assessment of Educational Progress* (NCES 2011–459). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Henningsen, M., & Stein, M. K. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for Research in Mathematics Education, 28,* 524–549. http://dx.doi.org/10.2307/749690

Hiebert, J. C., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester, Jr., (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. 1, pp. 371–404). New York, NY: Information Age Publishing.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42,* 371–406. http://dx.doi.org/10.3102/00028312042002371

Hirsch, B. J., & Rapkin, B. D. (1987). The transition to junior high school: A longitudinal study of self-esteem, psychological symptomatology, school life, and social support. *Child Development, 58,* 1235–1243. http://dx.doi.org/10.2307/1130617

Hitz, R., & Wright, D. (1988). Kindergarten issues: A practitioners' survey. *Principal, 67,* 28–30.

Huston, A. C., & Bentley, A. C. (2010). Human development in societal context. *Annual Review of Psychology, 61,* 411–437.

Huttenlocher, J., Vasilyeva, M., Waterfall, H. R., Vevea, J. L., & Hedges, L. V. (2007). The varieties of speech to young children. *Developmental Psychology, 43,* 1062–1083. http://dx.doi.org/10.1037/0012-1649.43.5.1062

Ing, M., Webb, N. M., Franke, M. L., Turrou, A. C., Wong, J., Shin, N., & Fernandez, C. H. (2014). *How student participation mediates the relationship between teacher practices and student achievement.* Paper presented at the annual meeting of the American Educational Research, Philadelphia, PA.

Jackson, K., & Wilson, J. (2012). Supporting African American students' learning of mathematics: A problem of practice. *Urban Education, 47,* 354–398. http://dx.doi.org/10.1177/0042085911429083

Jussim, L. (1989). Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology, 57,* 469–480. http://dx.doi.org/10.1037/0022-3514.57.3.469

Jussim, L., Eccles, J., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology, 28,* 281–388. http://dx.doi.org/10.1016/S0065-2601(08)60240-3

Klein, A., Starkey, P., & Wakeley, A. (2000). *Child Math Assessment: Preschool Battery (CMA).* Berkeley, CA: University of California.

Kozol, J. (1991). *Savage inequalities: Children in America's schools.* New York, NY: Crown Publishers.

Kuklinski, M. R., & Weinstein, R. S. (2001). Classroom and developmental differences in a path model of teacher expectancy effects. *Child Development, 72,* 1554–1578. http://dx.doi.org/10.1111/1467-8624.00365

Ladson-Billings, G. (1997). It doesn't add up: African American students' mathematics achievement. *Journal for Research in Mathematics Education, 28,* 697–708. http://dx.doi.org/10.2307/749638

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children enter school.* Washington, DC: Economic Policy Institute.

Loeb, S., & Bassok, D. (2008). Early childhood and the achievement gap. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of research in education policy finance and policy* (pp. 517–534). New York, NY: Routledge.

Lopez, N. (2002). *Hopeful girls, troubled boys: Race and gender disparity in urban education.* New York, NY: Routledge.

Lubienski, S. T. (2002). A closer look at Black-White mathematics gaps: Intersections of race and SES in NAEP achievement and instructional practices data. *Journal of Negro Education, 71,* 269–287. http://dx.doi.org/10.2307/3211180

Lubienski, S. T. (2006). Examining instruction, aehievement, and equity with NAEP mathematics data. *Education Policy Analysis Archives, 14,* 1–33.

Madon, S., Jussim, L., Keiper, S., Eccles, J., Smith, A., & Palumbo, P. (1998). The accuracy and power of sex, social class, and ethnic stereotypes: A naturalistic study in person perception. *Personality and Social Psychology Bulletin, 24,* 1304–1318. http://dx.doi.org/10.1177/01461672982412005

McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology, 46,* 235–261. http://dx.doi.org/10.1016/j.jsp.2007.05.001

Midgley, C., Feldlaufer, H., & Eccles, J. S. (1989). Change in teacher efficacy and student self-and task-related beliefs in mathematics during the transition to junior high school. *Journal of Educational Psychology, 81,* 247–258. http://dx.doi.org/10.1037/0022-0663.81.2.247

Muthén, L. K., & Muthén, B. O. (2013). Mplus version 7.1 [Computer software]. Los Angeles, CA: Author.

National Center for Education Statistics. (2013). *The nation's report card: Trends in academic progress 2012.* Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Council of Teachers of Mathematics. (1991). *1991–1992 handbook: NCTM goals, leaders, and positions.* Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

National Research Council. (2009). *Mathematics in early childhood: Learning paths toward excellence and equity*. Washington, DC: National Academy Press.

NICHD Early Child Care Research Network. (1996). Characteristics of infant child care: Factors contributing to positive caregiving. *Early Childhood Research Quarterly, 11,* 269–306.

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and ability grouping on opportunities to learn math and science*. Santa Monica, CA: RAND Corporation.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62,* 307–332. http://dx.doi.org/10.3102/00346543062003307

Peisner-Feinberg, E. S., & Burchinal, M. R. (1997). Relations between preschool children's child-care experiences and concurrent development: The Cost, Quality, and Outcomes Study. *Merrill-Palmer Quarterly, 43,* 451–477.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38,* 109–119.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15,* 209–233. http://dx.doi.org/10.1037/a0020141

Proctor, N. (1984). Towards a partnership with schools. *Journal of Education for Teaching, 10,* 219–232. http://dx.doi.org/10.1080/0260747840100303

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology, 76,* 85–97. http://dx.doi.org/10.1037/0022-0663.76.1.85

Reardon, S. F., & Galindo, C. (2006). *Patterns of Hispanic students' math and English literacy test scores in the early elementary grades*. National Task Force on Early Childhood Education for Hispanics.

Reardon, S. F., & Robinson, J. P. (2008). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 497–516). New York, NY: Routledge.

Reardon, S. F., Robinson-Cimpian, J. P., & Weathers, E. S. (2015). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In H. Ladd & M. Goertz (Eds.), *Handbook of research in education finance and policy* (2nd ed., pp. 491–509). Mahwah, NJ: Erlbaum.

Retelsdorf, J., Schwartz, K., & Asbrock, F. (2015). "Michael can't read!" Teachers' gender stereotypes and boys' reading self-concept. *Journal of Educational Psychology, 107,* 186–194. http://dx.doi.org/10.1037/a0037107

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review, 3,* 16–20. http://dx.doi.org/10.1007/BF02322211

Rubie-Davies, C. M. (2007). Classroom interactions: Exploring the practices of high- and low-expectation teachers. *British Journal of Educational Psychology, 77,* 289–306. http://dx.doi.org/10.1348/000709906X101601

Rubie-Davies, C. M., Peterson, E. R., Sibley, C. G., & Rosenthal, R. (2015). A teacher expectation intervention: Modeling the practices of high expectation teachers. *Contemporary Educational Psychology, 40,* 72–85. http://dx.doi.org/10.1016/j.cedpsych.2014.03.003

Rubie-Davies, C. M., Weinstein, R. S., Huang, F. L., Gregory, A., Cowan, P. A., & Cowan, C. P. (2014). Successive teacher expectation effects across the early school years. *Journal of Applied Developmental Psychology, 35,* 181–191. http://dx.doi.org/10.1016/j.appdev.2014.03.006

Rumberger, R., & Palardy, G. (2005). Does segregation still matter? The impact of student composition on academic achievement in high school. *Teachers College Record, 107,* 1999–2045.

Ruzek, E. A., Hafen, C. A., Allen, J. P., Gregory, A., Mikami, A. Y., & Pianta, R. C. (2016). How teacher emotional support motivates students: The mediating roles of perceived peer relatedness, autonomy support, and competence. *Learning and Instruction, 42,* 95–103. http://dx.doi.org/10.1016/j.learninstruc.2016.01.004

Saft, E. W., & Pianta, R. C. (2001). Teachers' perceptions of their relationships with students: Effects of child age, gender, and ethnicity of teachers and children. *School Psychology Quarterly, 16,* 125–141. http://dx.doi.org/10.1521/scpq.16.2.125.18698

Sarama, J., & Clements, D. H. (2013). Lessons learned in the implementation of the TRIAD scale-up model: Teaching early mathematics with trajectories and technologies. In T. G. Halle, A. J. Metz, & I. Martinez-Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. 173–191). Baltimore, MD: Brookes.

Sarama, J., Clements, D. H., Starkey, P., Klein, A., & Wakeley, A. (2008). Scaling up the implementation of a pre-kindergarten mathematics curriculumml: Teaching for understanding with trajectories and technologies. *Journal of Research on Educational Effectiveness, 1,* 89–119. http://dx.doi.org/10.1080/19345740801941332

Sarama, J., Clements, D. H., Wolfe, C. B., & Spitler, M. E. (2012). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies. *Journal of Research on Educational Effectiveness, 5,* 105–135. http://dx.doi.org/10.1080/19345747.2011.627980

Sarama, J., & DiBiase, A-M. (2004). The professional development challenge in preschool mathematics. In D. H. Clements, J. Sarama, & A.-M. DiBiase (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics education* (pp. 415–446). Mahwah, NJ: Erlbaum.

Schoen, H. L., Cebulla, K. J., Finn, K. F., & Fi, C. (2003). Teacher variables that relate to student achievement when using a standards-based curriculum. *Journal for Research in Mathematics Education, 34,* 228–259. http://dx.doi.org/10.2307/30034779

Selig, J. P., & Preacher, K. J. (2008). Monte Carlo method for assessing mediation: An interactive tool for creating confidence intervals for indirect effects [Computer software]. Retrieved from http://www.quantpsy.org

Seo, K., & Ginsburg, H. P. (2004). What is developmentally appropriate in early childhood mathematics education? Lessons from new research. In D. H. Clements & J. Sarama (Eds.), *Engaging young children in mathematics: Standards for early mathematics education* (pp. 91–104). Mahwah, NJ: Erlbaum.

Silver, E. A., & Stein, M. K. (1996). The Quasar Project: The "revolution of the possible" in mathematics instructional reform in urban middle schools. *Urban Education, 30,* 476–521. http://dx.doi.org/10.1177/0042085996030004006

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroomml: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85,* 571–581. http://dx.doi.org/10.1037/0022-0663.85.4.571

Sonnenschein, S., & Galindo, C. (2015). Race/ethnicity and early mathematics skills: Relations between home, classroom, and mathematics achievement. *The Journal of Educational Research, 108,* 261–277. http://dx.doi.org/10.1080/00220671.2014.880394

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629. http://dx.doi.org/10.1037/0003-066X.52.6.613

Stinson, D. W. (2006). African American male adolescents, schooling (and mathematics): Deficiency, rejection, and achievement. *Review of Educational Research, 76,* 477–506. http://dx.doi.org/10.3102/00346543076004477

Stipek, D. (1998). Motivation and instruction. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 85–113). New York, NY: Macmillan.

Stipek, D., & Byler, P. (2004). The early childhood classroom observation measure. *Early Childhood Research Quarterly, 19,* 375–397. http://dx.doi.org/10.1016/j.ecresq.2004.07.007

Stipek, D. J., Givvin, K. B., Salmon, J. M., & MacGyvers, V. L. (2001). Teachers' beliefs and practices related to mathematics instruction. *Teaching and Teacher Education, 17,* 213–226. http://dx.doi.org/10.1016/S0742-051X(00)00052-4

Stipek, D. J., & Ryan, R. H. (1997). Economically disadvantaged preschoolers: Ready to learn but further to go. *Developmental Psychology, 33,* 711–723. http://dx.doi.org/10.1037/0012-1649.33.4.711

Strutchens, M. E., & Silver, E. A. (2000). NAEP findings regarding race/ethnicity: Students' performance, school experiences, and attitudes and beliefs. In E. A. Silver, & P. A. Kenny (Eds.), *Results from the seventh mathematics assessment of the National Assessment of Educational Progress* (pp. 45–72). Reston, VA: National Council of Teachers of Mathematics.

Thompson, A. G. (1984). The relationship of teachers' conceptions of mathematics and mathematics teaching to instructional practice. *Educational Studies in Mathematics, 15,* 105–127. http://dx.doi.org/10.1007/BF00305892

Walshaw, M., & Anthony, G. (2008). The teacher's role in classroom discourse: A review of recent research into mathematics classrooms. *Review of Educational Research, 78,* 516–551. http://dx.doi.org/10.3102/0034654308320292

Weinstein, R. S., Soulé, C. R., Collins, F., Cone, J., Mehlhorn, M., & Simontacchi, K. (1991). Expectations and high school change: Teacher-researcher collaboration to prevent school failure. *American Journal of Community Psychology, 19,* 333–363.

Wenglinsky, H. (2004). Closing the racial achievement gap: The role of reforming instructional practices. *Education Policy Analysis Archives, 12*(64), 1–22.

West, C. K., & Anderson, T. H. (1976). The question of preponderant causation in teacher expectancy research. *Review of Educational Research, 46,* 613–630. http://dx.doi.org/10.3102/00346543046004613

Wigfield, A., Eccles, J. S., Mac Iver, D., Reuman, D. A., & Midgley, C. (1991). Transitions during early adolescence: Changes in children's domain-specific self-perceptions and general self-esteem across the transition to junior high school. *Developmental Psychology, 27,* 552–565. http://dx.doi.org/10.1037/0012-1649.27.4.552

Received February 8, 2016
Revision received October 12, 2016
Accepted October 19, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 6, 812–825

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000170

Classroom Stress Promotes Motivated Forgetting of Mathematics Knowledge

Gerardo Ramirez
University of California, Los Angeles

Ian M. McDonough
The University of Alabama

Ling Jin
University of California, Los Angeles

The ability to retain educationally relevant content in a readily accessible state in memory is critical for students at all stages in schooling. We hypothesized that a high degree of stress in mathematics courses can threaten students' mathematics self-concept and lead to a motivation to forget course content. We tested the aforementioned hypothesis by recruiting students from a college course on multivariate calculus. Students were asked to report their ongoing stress in the course. The forgetting rate was assessed by comparing students' final exam performance against their performance for a subset of the same final exam items 2 weeks later. We found that among students with a strong mathematics self-concept, a higher amount of ongoing weekly stress during the course was associated with increased forgetting of course content and a higher report of avoidant thinking about the course. Neither of these associations was found among students with a weaker mathematics self-concept. Our results provide evidence for a scientific account of the affective and motivational forces that shape why students forget educationally relevant content. We discuss the various educational practices that cue forgetting and make recommendations for reducing motivated forgetting in the classroom.

Keywords: classroom stress, memory, identity threat, mathematics performance, mathematics self-concept

A popular belief among students is that much of what is learned during class is forgotten soon after they are done with classes. Researchers commonly attribute forgetting to the disuse of the target information, interference from competing information, or the absence of the target information's retrieval cues. More recently, it has been argued that forgetting can also occur because of a deliberate process or motivation to exclude unwanted memories from our awareness (Anderson & Green, 2001). The motivated forgetting perspective broadly posits that internal drives can lead individuals to forget unpleasant memories that threaten the self. In the current study, we argue that one context in which people should be the most motivated to *remember* (i.e., the college classroom) might ironically create the motivation to *forget.* We reason that students for whom mathematics is a central part of their self-concept may be at risk for forgetting important classroom content when they undergo an unpleasant course experience.

Motivated Forgetting

Motivated forgetting is the process by which people have difficulty recalling information and memories for events that are unpleasant, painful, or generally threating to the self-perceptions that individuals strive to maintain (Ceci & Bruck, 1995; Thompson, Morton, & Fraser, 1997; Tajfel & Turner, 1986). By preventing the successful retrieval of these types of memories, motivated forgetting processes are believed to serve an adaptive function that helps preserve psychological well-being (DePrince et al., 2012). Research across clinical, social, and cognitive psychology provides strong evidence supporting this assertion. For instance, clinical and research accounts have reported that forgetting of traumatic information can be caused by memories of death, murder, and war (Arrigo & Pezdek, 1997; Belli, 2012; Pyszora, Barker, & Kopelman, 2003; Rivers, 1917), as well as childhood abuse at the hands of a trusted caregiver (Bowman, 1996; Herman & Schatzow, 1987).

Evidence of motivated forgetting has also been observed in more commonplace circumstances. Several studies investigating memory for historical passages show that people are less able to recall statements about historical atrocities when the perpetrators of those atrocities belong to the individual's cultural in-group (Imhoff & Banse, 2009; Rotella & Richeson, 2013). For instance, Sahdra and Ross (2007) found that Hindus who strongly identify with their in-group were the worst at recalling instances where Hindus engaged in aggression toward Sikhs (and vice versa). This

This article was published Online First February 20, 2017.

Gerardo Ramirez, Department of Psychology and Graduate School of Education & Information Studies, University of California, Los Angeles; Ian M. McDonough, Department of Psychology, The University of Alabama; Ling Jin, Graduate School of Education and Information Studies, University of California, Los Angeles.

We would like to thank David Taylor for his help with recruitment. This research was supported by Faculty Career Development Award to the PI from the University of California, Los Angeles.

Correspondence concerning this article should be addressed to Gerardo Ramirez, Department of Psychology and Graduate School of Education & Information Studies, University of California, Los Angeles, 1285 Franz Hall, CA 90095-1563. E-mail: gerardoramirez@ucla.edu

research suggests that one way individuals are able to maintain a self-serving history is by participating in a motivated forgetting process that reduces accessibility of historical information.

Individuals experience a similar motivated forgetting process when they receive individual level diagnostic history. Researchers investigating mnemic neglect ask participants to fill out personality surveys, which are then used to provide individuals with diagnostic feedback about their personality and behavior. This feedback varies as to whether it describes behaviors that are central or peripheral to the self, as well as whether the feedback is positive or negative. Oftentimes, when individuals are asked to recall previously given feedback, they are equally good at recalling both negative and positive feedback concerning behaviors peripheral to the self. However, when individuals are asked to recall feedback concerning behaviors central to the self, they are generally worse at recalling negative than positive feedback (Sedikides & Green, 2004). Work on mnemic neglect suggests that it is not simply that people are bad at recalling negative or unpleasant information, but rather that people show deficient recall (i.e., forgetting) of feedback that is threatening to a central part of the self.

More recent support for motivated forgetting has been informative in addressing a common criticism that can be raised about much of the aforementioned literature. For instance, one criticism of the motivated forgetting literature is that the threatening information that individuals experience might lead to deficient encoding of information rather than reductions in the retrievability of information after encoding (i.e., storage and retrieval). As an example of this deficient-encoding hypothesis, women show impaired note-taking activities following a manipulation that threatens their identity in science and mathematics (Appel, Kronberger, & Aronson, 2011). There are at least two studies that provide evidence that motivated forgetting can occur even when information was successfully encoded. These studies ask participants to encode neutral information prior to receiving information that is threatening to the self.

Shu, Gino, and Bazerman (2011) studied the ways in which people justify dishonest behaviors by forgetting moral rules. Shu et al. reasoned that when individuals commit dishonest behaviors that threaten their self-image as good and moral people (Aquino & Reed, 2002), it is advantageous that they strategically forget moral rules. To test this, researchers asked participants to memorize an ostensibly neutral "honor code," solve a problem-solving task, and then pay themselves for every problem they solved correctly. The authors found that participants who cheated by understating how much they actually paid themselves went on to remember less of the honor code during a memory test at the end of the experiment, compared with participants who did not cheat.

Another study asked whether people show reduced recall for neutral information linked to an identity threatened after the encoding phase. Dalton and Huang (2014) asked college students to memorize a set of neutral commercial advertisements that were either linked to their university (i.e., HKU students get 10% off) or not linked to their university (i.e., get 10% off). Following this neutral encoding task, participants either received threatening information (i.e., your institution is performing below other local universities) or nonthreatening information (i.e., your institution is on par with other local universities). Participants demonstrated poorer memory performance for the advertisements when they received the threatening relative to the nonthreatening information, but only if the advertisements were linked to their university. Dalton and Huang (2014) argued that threats to participants' university identity led to a motivation to forget even neutral information when linked to their threatened self-concept. Thus, in both studies participants encoded the information equally well, but motivational processes led to more forgetting of that information when individuals perceived it as a threat to the self.

Common Tenets in Motivated Forgetting

As reviewed previously, a diverse body of work provides ample evidence for motivated forgetting. Based on this evidence, we have identified two key tenets that inform our current research design and hypothesis. First, work on motivated forgetting suggests that people have some control over the process of forgetting. Basic memory research provides strong evidence that motives and directives can lead to real difficulty bringing memories to mind. For instance, Bjork and Bjork (1996) and others (e.g., Davis & Okada, 1971; MacLeod, 1998) maintain that explicitly directing individuals to forget words previously committed to memory can subsequently lead to an inability to recall this information (for review, see Bjork, Bjork, & Anderson, 1998). Participants in directed forgetting studies are not simply demonstrating demand effects; people continue to show poor recall even when they are provided a monetary incentive for successfully recalling the words they were previously asked to forget (MacLeod, 1999; Woodward & Bjork, 1971). Similarly, Anderson and Green (2001) demonstrate that the more times participants are instructed to inhibit particular words (using a think/no-think paradigm), the less likely they are to retrieve those words at a later recall test. Converging neuroimaging evidence also notes that this ability to control forgetting is associated with reduced activation in brain regions thought to be involved in memory retention (i.e., hippocampus), as well as increased activation in attentional control regions of the brain (i.e., the dorsolateral prefrontal cortex; Anderson et al., 2004; Benoit, Hulbert, Huddleston, & Anderson, 2015). There is clear support for the premise that individuals are capable of goal-directed forgetting via *external* directives using both the directed forgetting paradigm (Bjork, Bjork, & Anderson, 1998) and the think-no-think paradigm (Anderson & Green, 2001). Research on motivated forgetting simply extends this premise, indicating that certain situations create unpleasant experiences that trigger *internal* directives or motives to forget.

A second common tenet in the literature argues that the motivation to forget stems from the desire to protect one's self-concept. Self-concept refers to a predominantly positive mental representation of how people perceives themselves (Baumeister, 1998; Conway & Pleydell-Pearce, 2000; Gaertner, Sedikides, Vevea, & Iuzzini, 2002). The claim that threats to the self activate a type of psychological immune system with an impetus to protect one's self-concept goes back as far as Freud (1937), but is also a basic tenet in self-affirmation theory (Sherman & Cohen, 2006). Self-affirmation theory focuses on how individuals adapt and respond to experiences that threaten the self-concept. When situations create a threat to the self, individuals respond in a defensive manner and enlist a number of strategies that dismiss, deny, or avoid the threat in an effort to reaffirm the self. Some of these strategies involve making downward social comparisons against other individuals who are clearly inferior (Fein, Hoshino-Browne,

Davies, & Spencer, 2003), misidentifying or downplaying the importance of the domain they are performing in (Major, Spencer, Schmader, Wolfe, & Crocker, 1998), or distancing themselves from situations and people that confirm a threat (Goff, Steele, & Davies, 2008). We and others (Dalton & Huang, 2014; Green, Sedikides, & Gregg, 2008) borrow from self-affirmation theory to argue that memories that threaten a concept central to the self are likely to cue a pronounced motivation to forget.

We draw on the aforementioned tenets of the motivated forgetting literature to ask whether college students, who often are forced to contend with unpleasant course experiences, are at risk of forgetting important content once the class is over. We also address how students' own self-perceptions for the domain under study (in this case mathematics) moderate the effects of the unpleasant course experience to produce a motivation to forget.

Motivated Forgetting Within Education

One context where students are highly motivated to retain course content, but often report quickly forgetting, is the classroom (Bahrick, 1979). Curiously, there has been no work on motivated forgetting that focuses on real-world course content or fieldwork within the context of education. College classrooms, in particular, are an ideal place for studying motivated forgetting; students enter college with a strong academic self-concept yet are constantly challenged by a host of experiences in the classroom (Major et al., 1998; Marsh, 1991).

If motivated forgetting does occur within college courses, we reasoned that threats to students' academic self-concept most likely occur within challenging mathematics courses. Mathematics courses are often rated as the most difficult (Hoyt & Lee, 2002), receive the lowest course ratings (Centra, 2009), and commonly create aversive learning experiences for students across all stages in schooling (Hembree, 1990; Jackson & Leffingwell, 1999). In fact, many individuals suffer from a form of anxiety that is specific to mathematics (what is termed math anxiety). Math anxiety can emerge early in schooling (Ramirez, Chang, Maloney, Levine, & Beilock, 2016), stems from the negative attitudes and beliefs of parents as well as teachers (Beilock, Gunderson, Ramirez, & Levine, 2010; Maloney, Ramirez, Gunderson, Levine, & Beilock, 2015), and has even been associated with activation in brain regions associated with the visceral sensation of pain, threat, and vigilance (Lyons & Beilock, 2012; Pizzie & Krammer, 2016).

Stress and anxiety in regards to mathematics is a common problem even among college students who are quite competent and have strong perceptions about their mathematics ability. For instance, college students who can easily perform basic arithmetic operations and count dots are impaired under the duress of mathematics anxiety (Ashcraft & Kirk, 2001; Maloney, Ansari, & Fugelsang, 2011; Maloney, Risko, Ansari, & Fugelsang, 2010). In fact, international comparison studies reveal that students from countries with the highest levels of achievement (e.g., Singapore and Czech Republic) are most at risk for demonstrating declines in achievement because of mathematics anxiety (Organization for Economic Cooperation and Development, 2013). Mathematics is also a domain where negative ability stereotypes continue to exist even for highly competent college students. Women and underrepresented students in college must contend with negative ability stereotypes about their potential ability in mathematics, which can create additional stress and fear about confirming these existing negative stereotypes (Steele & Aronson, 1995). It is clear that stress and anxiety around mathematics are capable of derailing the educational outcomes of all students, even those who show high ability and great promise in mathematics.

If classroom experiences create threatening memories for students, this would suggest that the very same context where students are most motivated to remember course content might ironically create a powerful motivation for students to forget that course content instead. In fact, one study found that high-achieving women demonstrate more forgetting of algebra knowledge over time (Bahrick & Hall, 1991), which provides indirect evidence that populations who experience added stress in regards to mathematics are at risk for forgetting.

It is certainly the case that not all students who experience ongoing stress around their mathematics course may feel threatened by these experiences and therefore motivated to forget. Rather, students with high self-perceptions of mathematics ability (i.e., mathematics self-concept; Wigfield & Karpathian, 1991) should be the most vulnerable to forget. For students with a high mathematics self-concept, a high degree of ongoing course stress may undermine students' beliefs that their self-concept is adequately defined, stable, or internally consistent (Campbell et al., 1996). For example, experiencing social, financial, and work stress has been associated with a reduction in the extent to which individuals believe they have a clear idea of who they are (De Cremer & Sedikides, 2005; Lavallee & Campbell, 1995; Nezlek & Plesko, 2001; Ritchie, Sedikides, Wildschut, Arndt, & Gidron, 2011). Stressful classroom events may be threatening, in part, because they challenge one's daily assumptions and perceptions about their ability in mathematics. One way in which students might respond to such threats to the self is by creating a defensive response that seeks to minimize the accessibility of memories (i.e., motivated forgetting). Hence, students high in mathematics self-concept and whose course experience is characterized by a high degree of stress may come to interpret their stressful experiences as a threat to the assumptions they have about their abilities (e.g., "Mathematics is supposed to be something I am good at, so why am I feeling so stressed out by this class?"). If this is the case, then students with the highest self-concept for mathematics may be at risk for motivated forgetting. Thus, we ask:

> Research Question 1: To what extent does students' mathematics self-concept moderate the degree to which ongoing stress predicts forgetting of course material?

We focus on studying motivated forgetting at the conclusion of the course, when students may be most likely to deem the course content no longer relevant for retention (Bunce, VandenPlas, & Soulis, 2011; Khanna, Brack, & Finken, 2013). Hence, for the purposes of this study, we operationalized forgetting as the difference between students' performance on their final exam and their performance on a subset of the same final exam items given two weeks after the completion of the course. In this way, we tested whether stressful classroom experiences lead to motivated forgetting of classroom content after the course is over.

To the extent that we could find evidence for our first research question, we also must rule out several alternative interpretations. One possibility is that students with higher ongoing stress and

higher mathematics self-concept may perform very well on the original exam which, ironically, leaves them with more to lose on the follow-up exam two weeks later. A significant interaction between stress and mathematics self-concept on *original* exam performance would support this alternative. A second possibility is that a high degree of ongoing stress and high mathematics self-concept could lead to deficient encoding of the course content or exam preparation which would leave these students vulnerable to forgetting (i.e., the deficient-encoding hypothesis; Appel et al., 2011). Evidence in favor this account also would require that we demonstrate a relationship between ongoing stress and *original* final exam performance as a function of mathematics self-concept. To address this, we also asked:

> Research Question 2: To what extent does the combination of ongoing stress and having a high mathematics self-concept relate to performance on the original final exam?

While higher ongoing stress could lead students to differentially encode the course material, our intuition was that all students in the course would be equally motivated to effectively encode the material and perform well in this very important gateway course. As a preview, we carried out our study in an advanced multivariate calculus course that was geared for students interested in a science, technology, engineering or mathematics (STEM) career. We surmised that all students would correctly understand the need to maintain course content knowledge during their participation in the course to perform at a high level during exams. But once the class is over, a different story might arise. Students who go into their postcourse break (i.e., the summer) may see little need to reflect on and maintain memories that were threatening to their mathematics self-concept.

Additional alternative possibilities of the hypothesized motivated forgetting effect could be related to the amount of effort students put into the follow-up final exam given two weeks into the summer break. In general, we argue that individuals whose identity is threatened by their ongoing stress experience forget through a motivated process of inhibition that disrupts long-term memories related to the course content (i.e., a subtractive process that degrades retrieval strength; Erdelyi, 2006). However, students higher in ongoing stress and mathematics self-concept may simply not try as hard on the follow up assessment. Research on stereotype threat, for instance, documents that making stigmatized individuals aware of an existing negative stereotype can lead to reduced effort to perform (Schimel, Arndt, Banko, & Cook, 2004; Stone, 2002). Hence, we ask:

> Research Question 3: To what extent does students' self-reported effort during the follow-up exam vary as a function of ongoing stress and mathematics self-concept?

Lastly, students might avoid thoughts related to the course between the original exam and the follow-up final exam. This interpretation would implicate an avoidance strategy rather than inhibition as the cause of students' reduced ability to retrieve memories during the follow-up exam. Thought avoidance is a hallmark of stress and anxiety (Andrews et al., 2010) and a common response to identity threat (Brodish & Devine, 2009; Chalabaev, Sarranzinr, Stone, & Cury, 2008; Seibt & Förster, 2004). In fact, some work has determined that threats to the self can produce spontaneous performance-avoidance goals that are capable of reducing interest in the domain (Smith, Sansone, & White, 2007) and even lead to disidentification with the domain (Osborne, 1997; Osborne & Walker, 2006).

In terms of motivated forgetting, there is previous support that long-term memory recall is impaired by both inhibitory processes that suppress information at retrieval (Anderson et al., 2004; Butler & James, 2010), as well as avoidance processes that replace unpleasant memories with other thoughts (Hertel & Calcaterra, 2005). It is possible that, once the course is over, students whose mathematics self-concept was previously at risk might avoid thoughts related to their unpleasant course experience by occupying their mind with less threatening thoughts and memories (i.e., an additive process that adds "noise" to the signal; Erdelyi, 2006). Avoiding thoughts related to unpleasant memories also holds the potential to deliberately change the mental context altogether (Manning et al., 2016; Sahakyan & Kelley, 2002), which could impair students' ability to associate the previous course material with retrieval cues that are unique to the context after the course (i.e., the postcourse break). Hence, by measuring students' avoidance of course-related thoughts, we address the degree to which students were aware that they were actively making an effort to avoid thinking of the classroom material and whether this process accounts for forgetting. In our last question, we ask:

> Research Question 4: To what extent do students' reports of course-related thought avoidance account for the interaction between ongoing stress and mathematics self-concept on forgetting?

In summary, we attempt to provide the first evidence for motivated forgetting of educationally relevant materials within a natural classroom setting. This work is important in opening a new window into understanding the dynamic interpersonal processes that impact students' real-world retention of knowledge *after* the completion of a course. This new understanding, in turn, delivers important insights into the mechanisms for improving long-term retention of classroom content and supporting STEM students beyond their time in the classroom.

Method

Participants

We recruited students from an advanced multivariate calculus course at a large public university. The course is primarily offered to students who will enter STEM intensive fields that require comprehensive knowledge of advanced calculus. Four hundred students were enrolled in the course. Only $n = 185$ students gave their consent to participate in the study and filled out the initial web survey. Unfortunately, a large group of students eventually dropped the course, which reduced the study sample. The final sample of $n = 117$ consisted of students who completed both the original final exam as well as the follow-up exam. This sample of 117 students did not differ in gender, age, race, weekly stress, or mathematics self-concept from the 68 students who initially signed up in the study and dropped the course (all $ps > .05$). At the end of the study, all of the students who participated at any point in the study were paid \$30.

Materials and Procedure

The study consisted of four stages: an initial entry survey, weekly text messages, the original final exam, and a final assessment that included the follow-up exam.

Stage I: Initial Entry Survey. During the first week of the 10-week course, students were asked to complete the initial Web survey that asked students to answer a Demographic Survey and complete a Mathematics Self Concept Questionnaire. During this time students were also given instructions on how to respond to weekly text messages that they would receive (which we outline below).

Demographic information. Considering the large diversity that exists within the public university setting of our sample, we asked students to report their mother and father's level of education, completed from 1 (*less than high school*) to 7 (*graduate degree*), as well as their family's annual gross income from 1 (*less than $15,000*) to 7 (*$115,000 or more*). We created a composite measure of socioeconomic status (SES) by standardizing each of the three measures and averaging them together. We used this composite SES measure as a covariate in all of our analyses to account for individual differences in material, human and social resources of our sample of public university students. SES has also been previously shown to predict mathematics achievement (Dubow, Boxer, & Huesmann, 2009), knowledge retention across summer periods (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996), and chronic stress (Baum, Garofalo, & Yali, 1999; Cohen, Doyle, & Baum, 2006), making it a prime covariate to ensure that our results are not driven by differences in social capital that may exist among particular subgroups in our sample. We also asked students to report their intended college major and year in school in an effort to provide additional evidence that our sample was primarily interested in pursuing a STEM career.

Mathematics self-concept. We assessed students' mathematics self-concept using a six-item survey that broadly measures self-perceptions related to mathematics and were borrowed from previous identity threat research (e.g., Aronson et al., 1999; Beilock, Rydell, & McConnell, 2007; Markus, 1977; Smith, & White, 2001; Spencer, Steele, & Quinn, 1999). The six items were: (a) It is important to me that I am good at mathematics; (b) I am good at mathematics; (c) I feel that I am good at thinking analytically about mathematics; (d) I feel like I have a good understanding of mathematics concepts; (e) Compared with others, I feel that I perform well in mathematics; and (f) Compared with others, I feel I understand mathematics well. Participants responded to the six items using a 7-point Likert scale, from 1 (*strongly disagree*) to 7 (*strongly agree*). A reliability analysis revealed a Cronbach's alpha of .94 after we dropped one item (It is important to me that I am good at mathematics) that was reducing the reliability of the set of mathematics self-concept items. We subsequently averaged participants' responses on the remaining five items to create a mathematics self-concept index.

Ongoing stress. To get a measure of students' ongoing stress, we asked students to respond to the following question: *How stressed out do you feel in regards to the Mathematics course this week*? Students were required to enter with a numerical response from 1 (*not at all*) to 4 (*very much*). Participants entered their Week 1 responses during the initial entry survey and were informed that they needed to report their ongoing stress in regards to their mathematics course each week, which we outline in Stage II.

Stage II: Weekly text message. Students were sent a weekly text message for the remaining 9 weeks of the course that assessed their current mathematics stress level (i.e., ongoing stress). The same question ("*How stressed out do you feel in regards to the Mathematics course this week?*") and response scale (1–4) was used, as in the initial entry survey. Students received this text message on one of the two days they were scheduled to meet for class and were required to provide a response at some point during the day. We averaged across the 10 stress responses to create our measure of ongoing stress (Cronbach's alpha = .85).

Stage III: Final exam. During Week 11, the students who were still enrolled in the course were scheduled to take their final exam (referred to as the original final exam). The original final exam was cumulative and was built by the course instructor. The final exam was composed of 35 questions in a variety of formats (e.g., multiple choice, short answer, graph-equation matching, true-false). The combined set of items on the final exam showed weak internal consistency (Cronbach's alpha = .64). This low consistency between items was to be expected given that the exam covered a wide range of topics and was not meant to focus on a single, converging concept. A sample exam item was, "The linear approximation $L(x, y)$ of $f(x, y) = 5x - 100y$ at (0, 0) satisfies $L(x, y) = 5x - 100y$. Indicate whether the statement is true or false." Performance on the final exam served as our sole measure of actual course performance since we did not obtain permission to get the students' final grade. Students were allotted 2 hr to complete the original final exam.

Stage IV: Final assessment. To allow sufficient time for student forgetting, we waited two weeks after the end of the course before emailing the students to complete the final assessment of the study. The email contained a link that directed students to the final assessment that was administered using an online platform for data collection (Collector; https://github.com/gikeymarcia/Collector). Students began the final assessment by first reporting the extent to which they avoided thinking about their mathematics course using the following item: "*I have avoided thinking about my mathematics course since completing the class*" on a scale from 1 (*not at all true of me*) to 5 (*completely true of me*).

Students were then asked to complete the *follow-up exam* that consisted of the exact same questions they received for the original final exam in the course, with one exception: we only presented students with the multiple choice and true-false questions they previously received. The follow-up exam was composed of eight multiple-choice and eight true-false questions which together made up 40% of the original final exam. Our goal in only using the multiple choice and true-false problems was to limit the length of the follow-up exam to 30 min and to facilitate an objective scoring procedure that would not be dependent on teaching assistants to score the exam questions, which could introduce unnecessary variability. Our primary dependent variable was the difference between proportion correct on the original final exam and the follow-up exam (Follow-up Exam minus Original Final Exam) using the same questions across each test. After completing the follow-up exam, students were asked to report, "*How much effort did you put into the exam you just completed*" on a scale from 1 (*no effort*) to 5 (*all my effort*).

Results

Student Background and Scoring

The majority of our student sample was 18–19 years of age (M = 18.75, SD = .95). Within our sample, 89% of students reported an intention of going into a STEM career, 82% of students were in their first year of college, and 49% of our sample were female. The ethnic representation of our sample was Asian (53%), White (22%), Latino (17.4%), African American (3.5%), and other (4.4%). The frequencies of reported highest level of education for mothers and fathers, respectively, was less than high school (8.9% and 10.5%), high school (10.5% and 13.7%), at least 1 year college (8.9% and 10.5%), 2 years college (12.1% and 3.2%) 4 years of college (37.9% and 32.3%), some graduate training (3.2% for mothers), and graduate degree (18.5% and 29.8%). The frequencies of reported family income of our sample was less than $15,000 (7.9%), $15,000 to $34,999 (13.4%), $35,000 to $49,999 (14.2%), $50,000 to $74,999 (7.9%), $75,000 to $99,999 (18.1%), $100,000 to 114,999 (11.8%), $155,000 or more (26.8%).

Students responded to approximately 74.7% of the weekly text messages across the entire term and rated their average stress as moderate (M = 2.30, SD = .71). Looking at Figure 1, it is clear that student's ongoing stress varied dramatically across weeks with peaks near midterm and final exam. The mathematics self-concept average was above the center of the scale (M = 4.76, SD = 1.30). As a general rule of thumb, distributions with a skew between −1.0 and 1.0 are considered approximately symmetric (George & Mallery, 2010). The distribution of ongoing stress was slightly positively skewed (skewness score was .322) while the distribution of mathematics self-concept was slightly negatively skewed (−.460).

For the original final exam score, we analyzed only the specific items that were present in the follow-up Exam as a proportion of the total possible. Proportion correct on the original final exam (M = .76, SD = .13) was significantly higher than performance on the follow-up exam (M = .60, SD = .18; $t(117) = 9.65$, $p < .01$). Students showed an average reduction of .16 (SD = .18) proportion correct (a relative 21% drop in performance) from the original final exam to the follow-up exam. Table 1 presents the descriptive statistics and first-order correlations of our primary variables. Of note, students who reported higher mathematics self-concept also reported experiencing less ongoing stress in regards to their mathematics course, $r = -.37$, $p < .01$ and lower tendency to avoid thoughts related to their mathematics course, $r = -.23$, $p < .05$. These results suggest that higher mathematics self-concept is generally associated with a reduction in maladaptive course experiences and behaviors within our sample.

Before turning to our main analyses, we checked basic model assumptions for multivariate linear regression. As seen in Table 1, the strength of the correlation between ongoing stress and mathematics self-concept ($r = -.37$) did not warrant concern for multicollinearity. VIF values for our predictor variables were all between 1 and 2. We also plotted the model residuals in a histogram and found visual evidence for a normal distribution of residuals. A Durbin–Watson test for autocorrelation revealed a test statistic of 1.77, which is close to the value of 2 that is typically used as evidence that residuals are uncorrelated (i.e., independence of errors). A scatter plot of standardized predicted values against the standardized residuals demonstrated evidence that the residuals were constant. Admittedly, it would be more appropriate to analyze our ordinal data using nonparametric rather than the parametric tests. However, methodologists have argued that Likert items that are sums or averages across many items can be considered interval (see Norman, 2010, for an extensive discussion on this issue).

We addressed our main research questions by conducting a series of simultaneous regression models using PROCESS (Hayes, 2013). Main effect predictors were all mean centered. A listwise deletion procedure was used to deal with missing data across our model variables.[1] We present the full results for the regression models in Table 2.

> Research Question 1: To what extent does students' mathematics self-concept moderate the degree to which ongoing stress predicts forgetting of course material?

Our main moderation model (Model 1) evaluated whether average ongoing stress (the predictor variable) predicted forgetting rate (the primary outcome) and whether this relation was moderated by mathematics self-concept (the moderator variable) while controlling for SES (the control variable). The results for Model 1 revealed that the main effect of ongoing stress response and mathematics self-concept were not significant ($ps > .05$). However, consistent with our hypotheses, the interaction between students' ongoing stress response and mathematics self-concept was significant ($b = -.05$, $t = -2.62$, $p = .01$). A simple slopes analysis showed that at 1 standard deviation above the mean of mathematics self-concept, a higher ongoing stress response predicted student forgetting rate ($b = -.10$, 95% confidence interval [CI] [−.17, −.03]). In contrast, at both the mean and 1 standard deviation below the mean of mathematics self-concept, ongoing stress response did not relate to their forgetting rate ($ps > .05$; Figure 2. We also tested whether the aforementioned interaction would hold if not accounting for differences in SES. After removing SES from the model, the interaction between students' ongoing stress response and mathematics self-concept remained significant (see Model 2 in Table 2).

> Research Question 2: To what extent does the combination of ongoing stress and having a high mathematics self-concept relate to performance on the original final exam?

One alternative account of the results presented in Model 1 is that students with a higher mathematics self-concept and more ongoing stress may have performed the best on the original final exam and hence had more room to forget during the follow-up exam. We evaluated this alternative account by running the same set of predictors as Model 1, but with original final exam performance as our outcome (Model 3). The coefficient for the main effect of mathematics self-concept was significant ($b = .04$, $t = 3.79$, $p < .01$), revealing that greater mathematics self-concept was associated with greater performance on the original final exam. Critically, however, the main effect of ongoing stress and the

[1] There were 1–4 missing cases for each of our model predictors and outcome variables (<5% of the data for each respective variable). The results did not change when a mean imputation procedure was used instead to account for missing data.

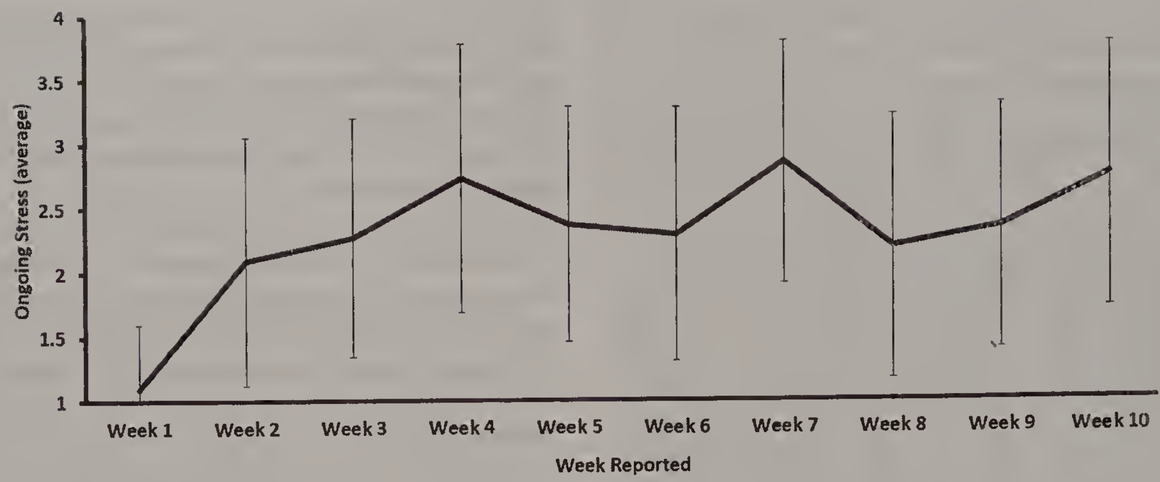


Figure 1. Ongoing stress across weeks. Error bars are *SD*s.

interaction between ongoing stress and mathematics self-concept were both not significant (*ps* $>$.05), arguing against the notion that students with a higher mathematics self-concept and more ongoing stress simply had more amount to forget or simply failed to encode the information adequately.

As stated earlier, our premise throughout the study is that students engage in a motivated forgetting process only once the class is over. Another way of providing evidence for this account is to predict students' follow-up exam performance when individuals are sharing the same original final exam pretest score (i.e., entering original final exam score as a covariate in the model rather than using a subtraction method). In this model (Model 4), we found no evidence for a main effect of mathematics self-concept ($p > .05$), but did find that the main effect of ongoing stress ($b = -.05$, $t = -2.35$, $p < .05$) and the interaction term were both significant ($b = -.05$, $t = -3.01$, $p < .01$). The results for Model 4 should not be surprising since a gain score approach (i.e., Follow up exam minus Original final exam) and an analysis of covariance (ANCOVA) approach are argued to be in general agreement (Maxwell & Delaney, 1999).

Research Question 3: To what extent does students' self-reported effort during the follow-up exam vary as a function of ongoing stress and mathematics self-concept?

We also evaluated whether our results might be explained by the amount of effort students were exerting in the follow-up exam. Students higher in mathematics self-concept may have been exerting less effort on the follow-up exam as a function of ongoing stress. We addressed this alternative account by using the same predictors in Model 1 to predict self-reported effort. Results for Model 5 showed that neither the main effects or interaction were significant (*ps* $>$ .05).

Research Question 4: To what extent do students' reports of course-related thought avoidance account for the interaction between ongoing stress and mathematics self-concept on forgetting?

To evaluate this question, we turn to students' self-reported tendency to avoid thinking about the course, which was measured during the follow-up assessment. In Model 6, we regressed students' self-reported tendency to avoid thinking about the course on ongoing stress, mathematics self-concept, and the interaction of these two factors (controlling for SES). The main effect of ongoing stress was not significant ($p > .05$), but we did find a significant main effect of mathematics self-concept ($b = -.22$, $t = -2.31$, $p < .05$) and a significant interaction between ongoing stress and mathematics self-concept ($b = .36$, $t = 2.82$, $p < .01$). For people at 1 *SD* above the mean in mathematics self-concept, greater ongoing stress was positively associated with students' self-reported tendency to avoid thinking about the course in the two weeks following the original final exam ($b = .66$ 95%, CI [.15, 1.16]). The simple slope between ongoing stress and avoiding thinking about the course content was not significant at the mean or 1 *SD* below the mean of mathematics self-concept (both $p >$ .05; Figure 3).

Table 1
Descriptive Statistics and Correlation Coefficients for Study Measures

Variable name	*M*	*SD*	1	2	3	4	5	6	7
(1) Forgetting score	−.16	.18	—						
(2) Original final exam	.76	.13	−.39**	—					
(3) Follow-up exam	.60	.18	.72**	.35**	—				
(4) Ongoing stress	2.30	.71	−.04	−.26**	−.23*	—			
(5) Math self-concept	4.76	1.30	−.15	.49**	.17	−.37**	—		
(6) Self-reported avoidance	3.07	1.27	.01	−.25**	−.18	.17	−.23*	—	
(7) Self-reported effort	2.57	.91	.09	.09	.16	−.01	.08	−.08	—
(8) Socioeconomic status	−.03	.83	−.07	.28**	.15	−.14	.13	−.23*	−.02

* $p < .05$. ** $p < .01$.

Table 2
Regression Models Predicting Forgetting Rate, Original Final Exam, Follow-Up Exam, and Student Self-Reported Avoidance

Variable	Forgetting rate		Original final exam	Follow-up exam	Self-report effort	Self-report avoidance	Forgetting rate
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Self-report avoidance							−.00
Original final exam				.44**			
Socioeconomic status	−.04*		.04*	−.02	−.04	−.13	−.04*
Math self-concept	−.02	.03	.04**	−.00	−.07	−.22*	−.02
Ongoing stress	.04	.04	−.02	−.05*	−.04	.19	−.04
Two way Interaction: Ongoing Stress × Math Self-Concept	−.05**	−.03*	−.01	−.05**	−.09	.36**	−.05*
R^2	.11	.08	.23	.24	.01	.16	.11
F	3.36	2.93	8.05	6.5	.37	5.02	2.67
F dfs	(4, 105)	(3, 106)	(4, 105)	(5, 104)	(4, 102)	(4, 104)	(5, 103)

Note. Values shown are unstandardized beta coefficients.
$^{*} p < .05$. $^{**} p < .01$.

Our last model tested whether the Ongoing Stress × Mathematics Self-Concept interaction would remain a significant predictor of forgetting rate, even after we added students' self-reported tendency to avoid thinking about the course as covariate. Our results for Model 7 revealed that the main effect of mathematics self-concept and ongoing stress response were not significant, but the interaction between students' ongoing stress response and mathematics self-concept remained significant ($b = -.05$, $t = -2.45$, $p = .016$). This finding suggests that avoidant thinking did not explain the forgetting rate results.

Discussion

Students are commonly tasked with learning important course content under the duress of classroom stress (Centra, 2009; Hoyt & Lee, 2002; Jackson & Leffingwell, 1999; Perry, 2004). We argue that ongoing stress throughout a mathematics course can create a threat for students with a high mathematics self-concept, which triggers a motivation to forget course content. We found support for this argument by demonstrating that higher reported ongoing stress was associated with more pronounced forgetting among students with a higher mathematics self-concept. For students with lower mathematics self-concept, higher ongoing stress for the mathematics course did not relate to the amount of forgotten mathematics content at the end of the course (i.e., during the postcourse break).

Our interpretation follows quite clearly from the common tenets in the motivated forgetting literature, which argue that unpleasant experiences that threaten the self can lead to a real difficulty in bringing unpleasant memories to mind. A good deal of behavioral and neuroimaging evidence also suggests that individuals are quite adept at suppressing unwanted memories (Anderson & Green, 2001; Anderson et al., 2004; Benoit, Hulbert, Huddleston, & Anderson, 2015; Bjork & Bjork, 1996; Bjork, Bjork, & Anderson, 1998; Manning et al., 2016). We borrow from an identity threat framework to propose that students with high self perceptions of mathematics ability are likely threatened by course experiences that challenge that perception, which leads to adaptive processes meant to reduce accessibility of unpleasant memories.

Critically, we argue that even though students may be motivated to forget while they are enrolled in their course, they avoid doing so since the material is still highly relevant to their performance and course grade. But once the course is over, those with a higher

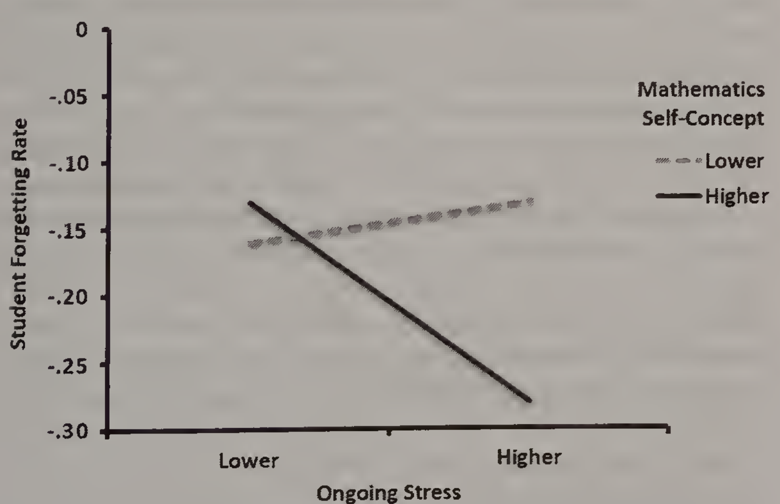


Figure 2. Interaction between ongoing stress and mathematics self-concept on forgetting rate (reported as a proportion). Ongoing stress and mathematics self-concept are plotted at 1 *SD* above and below the mean.

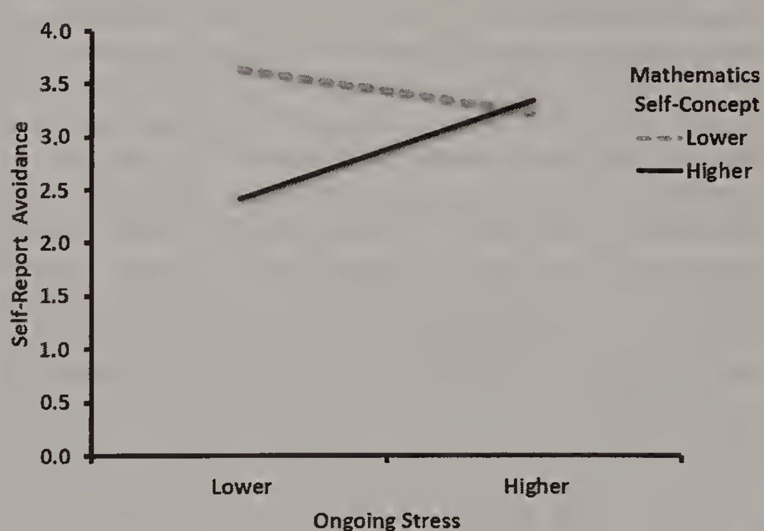


Figure 3. Interaction between ongoing stress and mathematics self-concept on self-reported avoidance. Ongoing stress and mathematics self-concept are plotted at 1 *SD* above and below the mean.

mathematics self-concept and who feel threatened by their previous course experience may feel unconstrained and give in to the motivation to forget as a means of protecting their perceptions of mathematics ability.

Of course, there are several alternative accounts for why we observed the pattern of results we report. One account is that students with a higher mathematics self-concept and more ongoing stress might simply have more room to forget. In other words, this subgroup of students might have performed very well on the original final exam, which ironically leaves them vulnerable to forget more on the follow-up exam. The opposite might also be true. Students higher in mathematics self-concept, who experiences account or the deficient-encoding account.

Another alternative account for our results is that students who are higher in mathematics self-concept and ongoing stress may have exerted reduced effort during the follow-up exam relative to their student peers (Schimel, Arndt, Banko, & Cook, 2004; Stone, 2002). We did not find evidence in support of this effort account. The results of Model 5 revealed that neither the main effects nor the interaction of ongoing stress and mathematics self-concept were significant predictors of self-reported effort during the follow-up exam.

We have primarily argued that motivated forgetting is driven by a suppression process that makes threatening information less retrievable. Another possibility, however, is that avoidance rather than suppression is the cause of students' reduced ability to retrieve memories. We examined this question by asking students how much they avoiding thinking about their mathematics course after the original final exam. We found that students with a higher mathematics self-concept who reported higher ongoing stress did, in fact, report that they avoided thinking about their mathematics course more than students with lower ongoing stress. The thought avoidance results suggest that students appear to have some awareness about their efforts to keep content outside of consciousness. Evidently students with a higher mathematics self-concept who do feel threatened (i.e., those who reported higher ongoing stress) not only forget more content, but also self-report avoidant thoughts about the course two weeks after the course is over.

Naturally one might wonder whether the forgetting results are explained by thought avoidance processes that, perhaps, make information less accessible. There are a number of possibilities by which thought avoidance could lead to greater forgetting. For instance, students who avoid thinking about (and retrieving) course relevant memories during their break could limit opportunities to associate their course memories with additional retrieval cues that are specific to the break context. Associating the course content solely with one context (i.e., the spring term) might reduce the match in retrieval cues that students draw upon during the follow-up exam (i.e., the summer term) and hence lead to difficulties in recall (Sahakyan & Kelley, 2002). It is also possible that avoiding thoughts related to the course could have been carried out by occupying one's mind with alternative thoughts and memories that disrupt long-term memory retrieval (Anderson et al., 2004; Benoit & Anderson, 2012; Butler & James, 2010).

While a number of reasons might explain why thought avoidance can lead to forgetting, we found that thought avoidance did not explain the forgetting effects in our study. In fact, we found that across the entire sample, students' self-reported tendency to avoid thinking about the course did not correlate with students' forgetting rate, $r = .01, p > .05$ (Table 1). Moreover, inspecting of Figures 2 and 3 reveal that the nature of the interaction between mathematics self-concept and ongoing stress also differed. Whereas forgetting rates were largest in the high mathematics self-concept/high stress group (Figure 2), avoidance ratings for this group of students were at a similarly high level as the low mathematics self-concept groups at both stress levels (Figure 3). The fact that these latter groups of students reported similar amounts of avoidance, but did not exhibit similar rates of forgetting suggest that a different mechanism is at play in the high mathematics self-concept/high stress group.

Limitations

A major limitation in this work is the correlational nature of our design that prevents us from making causal claims about the effects of ongoing stress on students' forgetting rates. The decision to conduct a correlational study was intentional, as examining motivated forgetting among STEM students allowed us to leverage the real-life variation in mathematics self-concept and weekly classroom stress that might predict forgetting.

Our focus in studying real-life course forgetting via exam performance also limited us in a number of other ways. For instance, the original final exam constructed by the course professor showed weak internal consistency. One possible explanation for the low internal consistency is that the course covered over 10 different topics (i.e., limits and continuity, calculus of vector value functions, gradient and directional derivatives, chain rule ad optimization, etc.) and each exam question incorporated 2 to 3 topics. This is not typical of other scales that focus on single psychological concepts. We were also limited by our decision to omit some final exam items during the follow up exam in an effort to keep the exam short and reduce student dropout. Similarly, we chose to only administer true-false and multiple-choice items, which limits the effects to recognition rather than recall processes. Prior work has found weaker forgetting effects on recognition compared with recall tests (Basden, Basden, & Gargano, 1993; Geiselman, Bjork, & Fishman, 1983; Gross, Barresi, & Smith, 1970; Wetzel, 1975), suggesting that our choice of materials on the follow-up exam may have made memory retrieval easier. It is also the case that readministering the 16 items from the original final exam on the follow-up exam builds in a test-retest effect that might underestimate the rate of forgetting. One way we could have avoided the inflation of scores because of previous testing is to form a new follow-up exam on the same concepts tested in the original final exam. However, following this line of logic makes the forgetting rate effects even more noteworthy because the recognition tests used for both the original final exam and the follow-up exam should have minimized the detection of such motivated forgetting effects.

Finally, we acknowledge that this study could have been strengthened if we had employed a separate measure of course performance outside of the final exam. However, we did not ask students for permission to obtain their course grades because we wanted to make the study minimally intrusive.

Educational Implications and Recommendations

Our results suggest that threatening classroom experiences may lead students to employ defense adaptations that unintentionally

impair memory for important course content. These defensive adaptations need to be addressed, possibly through interventions, to help students better cope with threats to their identity. For instance, it has been reported that threatening academic experiences can lead to a distortion in students' perceptions of previous mathematics performance and ability (Necka, Sokilowski, & Lyons, 2015), as well as disidentification and reduced interest in a domain of study (Steele, James, & Barnett, 2002). If threats to the self are indeed what underlie motivated forgetting in the classroom, then ensuring that students *leave* the classroom environment with a restored sense of self could help preserve the retention of classroom knowledge. Indeed, recent interest in using "wise-interventions" (Walton, 2014) to restore students' core social motives could be extended to help students at the end of the school year as well when students are likely to experience the highest levels of academic stress.

Of course, it is important to also give students the social-emotional skills to better understand and cope with ongoing stress as well. One promising method is to help students endorse a perspective that looks at stress (Crum, Salovey, & Achor, 2013), and failure more broadly (Haimovitz & Dweck, 2016), as an enhancing rather than as a threatening force. Prior work finds that reframing the physiological stress response as beneficial can lead to enhanced performance on standardized exams and school assignments (Jamieson, Mendes, & Nock, 2013). Students who approach classroom stress as a normal challenge that is a part of the learning process rather than a threat to their self-perception may have the appropriate appraisal perspective to minimize motivated forgetting in the classroom.

The work reported here also makes important recommendations for educators to be aware of the various educational practices that encourage forgetting. School systems and instructors widely engage in a host of classroom practices that create implicit cues to forget important classroom content, which may be exacerbated by the students' own motivation to forget.

At the level of the school system, the postcourse break (especially during the summer) is a time that provides a strong cue for students to forget what they have learned during the previous term. In fact, summer break has been identified as a period that is associated with profound forgetting because of a lack of educationally enriching activities (Cooper et al., 1996). However, the summer period may also lead to forgetting because it cues sharp event boundaries. Research on event cognition and memory (Kurby & Zacks, 2008) argues that memories for events that make up our daily life are segmented to have a beginning and an end, such that when people pass from one event to another, they forget more information than if they had not made such a shift (Radvansky & Copeland, 2006; Radvansky, Krawietz, & Tamplin, 2011). Summer break provides a concrete event boundary by which students might be cued to forget an unpleasant course experience. If this is the case, then extending mathematics activities into the summer could help to diffuse the event boundary between the school year and summer term and improve the retention of course content (Bahrick & Hall, 1991).

At the instructor level, pedagogical decisions impact rates of forgetting. For instance, classrooms that rely on a linear versus a spiral curriculum (Bunce, VandenPlas, & Soulis, 2011) and those in which instructors teach in a more traditional as opposed to inquiry-oriented manner (Kwon, Rasmussen, & Allen Keene, 2005) lead to greater difficulties in recalling STEM knowledge after the passage of time. Students also appear to search for reasons to retain content memories. A common student query is, "Why am I learning this?" which reflects the need to convey to students the importance of content. If answered appropriately, the students' need for relevance could go a long way in better promoting knowledge retention. For example, STEM textbooks that use applications to introduce and motivate a concept or skill result in greater retention of conceptual knowledge (Garner & Garner, 2001).

In addition, the presentation of blocked practice problems within mathematics textbooks (Rohrer & Taylor, 2007), as well as use of noncumulative examinations (Khanna, Brack, & Finken, 2013; Lawrence, 2013) are practices that also lead to greater forgetting. The aforementioned practices may lead to greater forgetting, in part, because they communicate that previously covered content no longer holds any relevance for future classroom performance.

Anecdotally, it seems like testing (cumulative or not) is a common source of ongoing stress for many students that could tempt educators to reduce testing altogether. Yet distributed testing actually seems to reduce testing-related stress (Agarwal, D'Antonio, Roediger, McDermott, & McDaniel, 2014; Crooks, 1988; Dempster, 1992; Dustin, 1971; Szpunar, Khan, & Schacter, 2013) and promote long-term knowledge retention (McDaniel, Anderson, Derbish, & Morrisette, 2007; Roediger & Karpicke, 2006). Hence, when tests are used as a learning tool, more rather than less testing may give teachers one promising avenue by which to improve knowledge retention and reduce ongoing stress that contributes to motivated forgetting.

At the student level, we suspect that particular study practices such as taking photographs of classroom PowerPoint slides (Henkel, 2014) or taking long-hand notes on a laptop during lectures (Mueller & Oppenheimer, 2014) have the potential to encourage forgetting as well. These practices lead students to expect that they will have future access to the information and do not need to maintain a strong memory representation (Sparrow, Liu, & Wegner, 2011).

Our work contributes to ongoing discussions around whether the forgetting process that underlies motivated forgetting is largely carried out as a result of students' conscious efforts to suppress threatening information or whether this process occurs below conscious awareness. Indeed, there is a lively ongoing debate (Brewin & Andrews, 2014; Handy, 2015; Patihis, Ho, Tingen, Lilienfeld, & Loftus, 2014) about whether motivated forgetting is best explained by controlled conscious processes (i.e., suppression) or processes outside of conscious awareness (i.e., repression; Epstein, 1994; Freud, 1937), with others suggesting that repression and suppression are largely the same thing (Erdelyi, 2006). Our forgetting results highlight that students may forget the mathematics course content despite being aware that this content will be relevant in their future studies. The thought avoidance results suggest that this motivated forgetting process is also paired with a conscious effort to avoid bringing threatening content to mind. We hope that the work reported here will stimulate this ongoing discussion on whether motivated forgetting is a process that is largely outside of conscious awareness, by providing an example of motivated forgetting within a natural classroom field setting.

Conclusion

The ability to retain learned material in a course is critical to excel in related courses and to succeed in a work environment that depends on those learned skills and knowledge. We provide the first evidence of a self-directed motivation that jeopardizes long-term retention of course material in a real-world educational context. Students with a high self-concept related to the course topic and ongoing course-related stress are at the greatest risk for forgetting that material. Future work aimed at supporting long-term retention of course material is critical for this group of students who, ironically, are most in need of retaining the material for future STEM courses and careers.

References

Agarwal, P. K., D'Antonio, L., Roediger, H. L., III, McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory & Cognition, 3,* 131–139. http://dx.doi.org/10.1016/j.jarmac.2014.07.002

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature, 410,* 366–369. http://dx.doi.org/10.1038/35066572

Anderson, M. C., Ochsner, K. N., Kuhl, B., Cooper, J., Robertson, E., Gabrieli, S. W., . . . Gabrieli, J. D. (2004). Neural systems underlying the suppression of unwanted memories. *Science, 303,* 232–235. http://dx.doi.org/10.1126/science.1089504

Andrews, G., Hobbs, M. J., Borkovec, T. D., Beesdo, K., Craske, M. G., Heimberg, R. G., . . . Stanley, M. A. (2010). Generalized worry disorder: A review of *DSM–IV* generalized anxiety disorder and options for DSM-V. *Depression and Anxiety, 27,* 134–147. http://dx.doi.org/10.1002/da.20658

Appel, M., Kronberger, N., & Aronson, J. (2011). Stereotype threat impedes ability building: Effects on the test preparation of women in science and technology. *European Journal of Social Psychology, 41,* 904–913. http://dx.doi.org/10.1002/ejsp.835

Aquino, K., & Reed, A., II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology, 83,* 1423–1440. http://dx.doi.org/10.1037/0022-3514.83.6.1423

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When White men can't do mathematics: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35,* 29–46. http://dx.doi.org/10.1006/jesp.1998.1371

Arrigo, J. M., & Pezdek, K. (1997). Lessons from the study of psychogenic amnesia. *Current Directions in Psychological Science, 6,* 148–152. http://dx.doi.org/10.1111/1467-8721.ep10772916

Ashcraft, M., & Kirk, E. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General, 1302,* 224–237.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108,* 296–308. http://dx.doi.org/10.1037/0096-3445.108.3.296

Bahrick, H. P., & Hall, L. K. (1991). Lifetime maintenance of high school mathematics content. *Journal of Experimental Psychology: General, 120,* 20–33. http://dx.doi.org/10.1037/0096-3445.120.1.20

Basden, B. H., Basden, D. R., & Gargano, G. J. (1993). Directed forgetting in implicit and explicit memory tests: A comparison of methods. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 603–616. http://dx.doi.org/10.1037/0278-7393.19.3.603

Baum, A., Garofalo, J. P., & Yali, A. M. (1999). Socioeconomic status and chronic stress. Does stress account for SES effects on health? *Annals of the New York Academy of Sciences, 896,* 131–144. http://dx.doi.org/10.1111/j.1749-6632.1999.tb08111.x

Baumeister, R. F. (1998). The self. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., pp. 680–740). New York, NY: McGraw-Hill.

Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences of the United States of America, 107,* 1860–1863. http://dx.doi.org/10.1073/pnas.0910967107

Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General, 136,* 256–276. http://dx.doi.org/10.1037/0096-3445.136.2.256

Belli, R. F. (2012). Introduction: In the aftermathematics of the so-called memory wars. In R. F. Belli (Ed.), *True and false recovered memories: Toward a reconciliation of the debate* (pp. 1–13). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-1195-6_1

Benoit, R. G., & Anderson, M. C. (2012). Opposing mechanisms support the voluntary forgetting of unwanted memories. *Neuron, 76,* 450–460. http://dx.doi.org/10.1016/j.neuron.2012.07.025

Benoit, R. G., Hulbert, J. C., Huddleston, E., & Anderson, M. C. (2015). Adaptive top-down suppression of hippocampal activity and the purging of intrusive memories from consciousness. *Journal of Cognitive Neuroscience, 27,* 96–111. http://dx.doi.org/10.1162/jocn_a_00696

Bjork, E. L., & Bjork, R. A. (1996). Continuing influences of to-be-forgotten information. *Consciousness and Cognition, 5,* 176–196. http://dx.doi.org/10.1006/ccog.1996.0011

Bjork, E. L., Bjork, R. A., & Anderson, M. C. (1998). Varieties of goal-directed forgetting. In J. M. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 103–137). Mahwah, NJ: Erlbaum.

Bowman, E. S. (1996). Delayed memories of child abuse: Part I: An overview of research findings on forgetting, remembering, and corroborating trauma. *Dissociation: The Official Journal of the International Society for the Study of Multiple Personality & Dissociation, 9,* 221–231.

Brewin, C. R., & Andrews, B. (2014). Why it is scientifically respectable to believe in repression: A response to Patihis, Ho, Tingen, Lilienfeld, and Loftus (2014). *Psychological Science, 25,* 1964–1966. http://dx.doi.org/10.1177/0956797614541856

Brodish, A. B., & Devine, P. G. (2009). The role of performance–avoidance goals and worry in mediating the relationship between stereotype threat and performance. *Journal of Experimental Social Psychology, 45,* 180–185. http://dx.doi.org/10.1016/j.jesp.2008.08.005

Bunce, D. M., VandenPlas, J. R., & Soulis, C. (2011). Decay of student knowledge in chemistry. *Journal of Chemical Education, 88,* 1231–1237. http://dx.doi.org/10.1021/ed100683h

Butler, A. J., & James, K. H. (2010). The neural correlates of attempting to suppress negative versus neutral memories. *Cognitive, Affective & Behavioral Neuroscience, 10,* 182–194. http://dx.doi.org/10.3758/CABN.10.2.182

Campbell, J. D., Trapnell, P. D., Heine, S. J., Katz, I. M., Lavallee, L. F., & Lehman, D. R. (1996). Self-concept clarity: Measurement, personality correlates, and cultural boundaries. *Journal of Personality and Social Psychology, 70,* 141–156.

Ceci, S. J., & Bruck, M. (1995). *Jeopardy in the courtroom: A scientific analysis of children's testimony.* Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10180-000

Centra, J. A. (2009). *Differences in responses to the student instructional report: Is it bias?* Princeton, NJ: Educational Testing Service.

Chalabaev, A., Sarrazinr, P., Stone, J., & Cury, F. (2008). Do achievement goals mediate stereotype threat? An investigation on females' soccer performance. *Journal of Sport & Exercise Psychology, 30,* 143–158. http://dx.doi.org/10.1123/jsep.30.2.143

Cohen, S., Doyle, W. J., & Baum, A. (2006). Socioeconomic status is associated with stress hormones. *Psychosomatic Medicine, 68,* 414–420. http://dx.doi.org/10.1097/01.psy.0000221236.37158.b9

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review, 107,* 261–288. http://dx.doi.org/10.1037/0033-295X.107.2.261

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66,* 227–268. http://dx.doi.org/10.3102/00346543066003227

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58,* 438–481. http://dx.doi.org/10.3102/00346543058004438

Crum, A., Salovey, P., & Achor, S. (2013). The role of mindsets in determining the stress response. *Journal of Personality and Sociual Psychology, 104,* 716–733.

Dalton, A. N., & Huang, L. (2014). Motivated forgetting in response to social identity threat. *Journal of Consumer Research, 40,* 1017–1038. http://dx.doi.org/10.1086/674198

Davis, J. C., & Okada, R. (1971). Recognition and recall of positively forgotten items. *Journal of Experimental Psychoiogy, 89,* 181–186. http://dx.doi.org/10.1037/h0031183

De Cremer, D., & Sedikides, C. (2005). Self-uncertainty and responsiveness to procedural justice. *Journal of Experimental Social Psychology, 41,* 157–173. http://dx.doi.org/10.1016/j.jesp.2004.06.010

Dempster, F. N. (1992). Using tests to promote learning: A neglected classroom resource. *Journal of Research & Development in Education, 25,* 213–217.

DePrince, A. P., Brown, L. S., Cheit, R. E., Freyd, J. J., Gold, S. N., Pezdek, K., & Quina, K. (2012). Motivated forgetting and misremembering: Perspectives from betrayal trauma theory. In *True and false recovered memories* (pp. 193–242). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-1195-6_7

Dubow, E. F., Boxer, P., & Huesmann, L. R. (2009). Long-term effects of parents' education on children's educational and occupational success: Mediation by family interactions, child aggression, and teenage aspirations. *Merrill-Palmer Quarterly, 55,* 224–249. http://dx.doi.org/10.1353/mpq.0.0030

Dustin, D. S. (1971). Some effects of exam frequency. *The Psychological Record, 2,* 409–414.

Epstein, S. (1994). Integration of the cognitive and psychodynamic unconscious. *American Psychologist, 49,* 709–724.

Erdelyi, M. H. (2006). The unified theory of repression. *Behavioral and Brain Sciences, 29,* 499–511. http://dx.doi.org/10.1017/S0140525X06009113

Fein, S., Hoshino-Browne, E., Davies, P. G., & Spencer, S. J. (2003). Self-image maintenance goals and sociocultural norms in motivated social perception. In S. J. Spencer, S. Fein, M. P. Zanna, & J. Olson (Eds.), *Motivated social perception: The Ontario symposium* (Vol. 9, pp. 21–44). Mahwah, NJ: Erlbaum.

Freud, A. (1937). *The ego and the mechanisms of defense.* London, England: Hogarth Press and Institute of Psycho-Analysis.

Gaertner, L., Sedikides, C., Vevea, J. L., & Iuzzini, J. (2002). The "I," the "we," and the "when": A meta-analysis of motivational primacy in self-definition. *Journal of Personality and Social Psychology, 83,* 574–591. http://dx.doi.org/10.1037/0022-3514.83.3.574

Garner, B. E., & Garnder, L. E. (2001). Retention of concepts and skills in traditional and reformed applied calculus. *Mathematics Education Research Journal, 13,* 165–184. http://dx.doi.org/10.1007/BF03217107

Geiselman, R. E., Bjork, R. A., & Fishman, D. L. (1983). Disrupted retrieval in directed forgetting: A link with posthypnotic amnesia. *Journal of Experimental Psychology: General, 112,* 58–72. http://dx.doi.org/10.1037/0096-3445.112.1.58

George, D., & Mallery, M. (2010). *SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 update* (10th ed.). Boston, MA: Pearson.

Goff, P. A., Steele, C. M., & Davies, P. G. (2008). The space between us: Stereotype threat and distance in interracial contexts. *Journal of Personality and Social Psychology, 94,* 91–107. http://dx.doi.org/10.1037/0022-3514.94.1.91

Green, J. D., Sedikides, C., & Gregg, A. P. (2008). Forgotten but not gone: The recall and recognition of self-threatening memories. *Journal of Experimental Social Psychology, 44,* 547–561.

Gross, A. E., Barresi, J., & Smith, E. E. (1970). Voluntary forgetting of a shared memory load. *Psychonomic Science, 20,* 73–75. http://dx.doi.org/10.3758/BF03335607

Haimovitz, K., & Dweck, C. S. (2016). What predicts children's fixed and growth intelligence mind-sets? Not their parents' views of intelligence but their parents' views of failure. *Psychological Science, 27,* 859–869. http://dx.doi.org/10.1177/0956797616639727

Handy, J. D. (2015). *The continued march towards ecological validity in laboratory studies of blocked and recovered memories* (Doctoral dissertation). Retrieved from http://hdl.handle.net/1969.1/155158

Hayes, A. F. (2013). *An introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* New York, NY: Guilford Press.

Hembree, R. (1990). The nature, effects, and relief of mathematicsematics anxiety. *Journal for Research in Mathematicsematics Education, 21,* 33–46. http://dx.doi.org/10.2307/749455

Henkel, L. A. (2014). Point and shoot memories: The influence of taking photos on memory for a museum tour. *Psychological Science, 25,* 396–402.

Herman, J. L., & Schatzow, E. (1987). Recovery and verification of memories of childhood sexual trauma. *Psychoanalytic Psychology, 4,* 1–14. http://dx.doi.org/10.1037/h0079126

Hertel, P. T., & Calcaterra, G. (2005). Intentional forgetting benefits from thought substitution. *Psychonomic Bulletin & Review, 12,* 484–489. http://dx.doi.org/10.3758/BF03193792

Hoyt, D. P., & Lee, E. J. (2002). Teaching styles and learning outcomes. *IDEA Research Report.* Retrieved from http://www.idea.ksu.edu/reports/research4.pdf

Imhoff, R., & Banse, R. (2009). Ongoing victim suffering increases prejudice: The case of secondary anti-semitism. *Psychological Science, 20,* 1443–1447. http://dx.doi.org/10.1111/j.1467-9280.2009.02457.x

Jackson, C. D., & Leffingwell, R. J. (1999). The role of instructors in creating mathematics anxiety in students from kindergarten through college. *The Mathematicsematics Teacher, 92,* 583–586.

Jamieson, J. P., Mendes, W. B., & Nock, M. K. (2013). Improving acute stress responses: The power of reappraisal. *Current Directions in Psychological Science, 22,* 51–56. http://dx.doi.org/10.1177/0963721412461500

Khanna, M. M., Brack, A. S. B., & Finken, L. L. (2013). Short- and long-term effects of cumulative finals on student learning. *Teaching of Psychology, 40,* 175–182. http://dx.doi.org/10.1177/0098628313487458

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences, 12,* 72–79. http://dx.doi.org/10.1016/j.tics.2007.11.004

Kwon, O. H., Rasmussen, C., & Allen Keene, K. (2005). Students' retention of mathematical knowledge and skills in differential equations. *School Science and Mathematics, 105,* 227–239. http://dx.doi.org/10.1111/j.1949-8594.2005.tb18163.x

Lavallee, L. F., & Campbell, J. D. (1995). Impact of personal goals on self-regulation processes elicited by daily negative events. *Journal of Personality and Social Psychology, 69,* 341–352. http://dx.doi.org/10.1037/0022-3514.69.2.341

Lawrence, N. K. (2013). Cumulative exams in the introductory psychology course. *Teaching of Psychology, 40,* 15–19. http://dx.doi.org/10.1177/0098628312465858

Lyons, I. M., & Beilock, S. L. (2012). When math hurts: Math anxiety predicts pain network activation in anticipation of doing math. *PLOS ONE, 791,* e48076.

MacLeod, C. M. (1998). Directed forgetting. In J. M. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 1–57). Mahwah, NJ: Erlbaum.

MacLeod, C. M. (1999). The item and list methods of directed forgetting: Test differences and the role of demand characteristics. *Psychonomic Bulletin & Review, 6,* 123–129. http://dx.doi.org/10.3758/BF03210819

Major, B., Spencer, S., Schmader, T., Wolfe, C., & Crocker, J. (1998). Coping with negative stereotypes about intellectual performance: The role of psychological disengagement. *Personality and Social Psychology Bulletin, 24,* 34–50. http://dx.doi.org/10.1177/0146167298241003

Maloney, E. A., Ansari, D., & Fugelsang, J. A. (2011). The effect of mathematics anxiety on the processing of numerical magnitude. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 64,* 10–16. http://dx.doi.org/10.1080/17470218.2010.533278

Maloney, E. A., Ramirez, G., Gunderson, E. A., Levine, S. C., & Beilock, S. L. (2015). Intergenerational effects of low math achievement and high math anxiety. *Psychological Science, 26,* 1480–1488. http://dx.doi.org/10.1177/0956797615592630

Maloney, E. A., Risko, E. F., Ansari, D., & Fugelsang, J. (2010). Mathematics anxiety affects counting but not subitizing during visual enumeration. *Cognition, 114,* 293–297. http://dx.doi.org/10.1016/j.cognition.2009.09.013

Manning, J. R., Hulbert, J. C., Williams, J., Piloto, L., Sahakyan, L., & Norman, K. A. (2016). A neural signature of contextually mediated intentional forgetting. *Psychonomic Bulletin & Review, 23,* 1534–1542.

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology, 35,* 63–78. http://dx.doi.org/10.1037/0022-3514.35.2.63

Marsh, H. W. (1991). Failure of high-ability high schools to deliver academic benefits commensurate with their students' ability levels. *American Educational Research Journal, 28,* 445–480. http://dx.doi.org/10.3102/00028312028002445

Maxwell, S. E., & Delaney, H. D. (1999). *Designing experiments and analyzing data: A model comparison perspective.* Mahwah, NJ: Erlbaum.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19,* 494–513. http://dx.doi.org/10.1080/09541440701326154

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science, 25,* 1159–1168.

Necka, E. A., Sokolowski, H. M., & Lyons, I. M. (2015). The role of self-math overlap in understanding math anxiety and the relation between math anxiety and performance. *Frontiers in Psychology, 6,* 1543.

Nezlek, J. B., & Plesko, R. M. (2001). Day-to-day relationships among self-concept clarity, self-esteem, daily events, and mood. *Personality and Social Psychology Bulletin, 27,* 201–211. http://dx.doi.org/10.1177/0146167201272006

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15,* 624–632.

Organization for Economic Cooperation and Development. (2013). PISA 2012 results: Ready to learn: Students' engagement, drive and self-beliefs (Volume III). OECD Publishing. http://dx.doi.org/10.1787/9789264201170-en

Osborne, J. W. (1997). Race and academic disidentification. *Journal of Educational Psychology, 89,* 728–735. http://dx.doi.org/10.1037/0022-0663.89.4.728

Osborne, J. W., & Walker, C. (2006). Stereotype threat, identification with academics, and withdrawal from school: Why the most successful students of colour might be most likely to withdraw. *Educational Psychology, 26,* 563–577. http://dx.doi.org/10.1080/01443410500342518

Patihis, L., Ho, L. Y., Tingen, I. W., Lilienfeld, S. O., & Loftus, E. F. (2014). Are the "memory wars" over? A scientist-practitioner gap in beliefs about repressed memory. *Psychological Science, 25,* 519–530. http://dx.doi.org/10.1177/0956797613510718

Perry, A. B. (2004). Decreasing mathematics anxiety in college students. *College Student Journal, 38,* 321–324.

Pizzie, R. G., & Krammer, D. J. M. (2016, May). *Anxious attention: Math anxiety predicts amygdala reactivity to mathematical stimuli.* Poster presented at the International Meeting of Psychonomics Society, Granada, Spain.

Pyszora, N., Barker, A., & Kopelman, M. (2003). Amnesia for criminal offences: A study of life sentence prisoners. *Journal of Forensic Psychiatry & Psychology, 14,* 475–490. http://dx.doi.org/10.1080/14789940310001599785

Radvansky, G. A., & Copeland, D. E. (2006). Walking through doorways causes forgetting: Situation models and experienced space. *Memory & Cognition, 34,* 1150–1156. http://dx.doi.org/10.3758/BF03193261

Radvansky, G. A., Krawietz, S. A., & Tamplin, A. K. (2011). Walking through doorways causes forgetting: Further explorations. *The Quarterly Journal of Experimental Psychology, 64,* 1632–1645. http://dx.doi.org/10.1080/17470218.2011.571267

Ramirez, G., Chang, H., Maloney, E. A., Levine, S. C., & Beilock, S. L. (2016). On the relationship between math anxiety and math achievement in early elementary school: The role of problem solving strategies. *Journal of Experimental Child Psychology, 141,* 83–100. http://dx.doi.org/10.1016/j.jecp.2015.07.014

Ritchie, T. D., Sedikides, C., Wildschut, T., Arndt, J., & Gidron, Y. (2011). Self-concept clarity mediates the relation between stress and subjective well-being. *Self and Identity, 10,* 493–508. http://dx.doi.org/10.1080/15298868.2010.493066

Rivers, W. H. R. (1917). Freud's psychology of the unconscious. *The Lancet, 189,* 912–914. http://dx.doi.org/10.1016/S0140-6736(00)44819-1

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. http://dx.doi.org/10.1111/j.1745-6916.2006.00012.x

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35,* 481–498. http://dx.doi.org/10.1007/s11251-007-9015-8

Rotella, K. N., & Richeson, J. A. (2013). Motivated to "forget": The effects of in-group wrongdoing on memory and collective guilt. *Social Psychological and Personality Science, 4,* 730–737. http://dx.doi.org/10.1177/1948550613482986

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 1064–1072. http://dx.doi.org/10.1037/0278-7393.28.6.1064

Sahdra, B., & Ross, M. (2007). Group identification and historical memory. *Personality and Social Psychology Bulletin, 33,* 384–395. http://dx.doi.org/10.1177/0146167206296103

Schimel, J., Arndt, J., Banko, K. M., & Cook, A. (2004). Not all self-affirmations were created equal: The cognitive and social benefits of affirming the intrinsic (vs. extrinsic) self. *Social Cognition, 22,* 75–99. http://dx.doi.org/10.1521/soco.22.1.75.30984

Sedikides, C., & Green, J. D. (2004). What I don't recall can't hurt me: Information negativity versus information inconsistency as determinants of memorial self-defense. *Social Cognition, 22,* 4–29. http://dx.doi.org/10.1521/soco.22.1.4.30987

Seibt, B., & Förster, J. (2004). Stereotype threat and performance: How self-stereotypes influence processing by inducing regulatory foci. *Journal of Personality and Social Psychology, 87,* 38–56. http://dx.doi.org/10.1037/0022-3514.87.1.38

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. *Advances in Experimental Social Psychology, 38,* 183–242. http://dx.doi.org/10.1016/S0065-2601(06)38004-5

Shu, L. L., Gino, F., & Bazerman, M. H. (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and Social Psychology Bulletin, 37,* 330–349. http://dx.doi.org/10.1177/0146167211398138

Smith, J. L., Sansone, C., & White, P. H. (2007). The stereotyped task engagement process: The role of interest and achievement motivation. *Journal of Educational Psychology, 99,* 99–114. http://dx.doi.org/10.1037/0022-0663.99.1.99

Smith, J. L., & White, P. H. (2001). Development of the domain identification measure: A tool for investigating stereotype threat effects. *Educational and Psychological Measurement, 61,* 1040–1057. http://dx.doi.org/10.1177/00131640121971635

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333,* 776–778. http://dx.doi.org/10.1126/science.1207745

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's mathematics performance. *Journal of Experimental Social Psychology, 35,* 4–28. http://dx.doi.org/10.1006/jesp.1998.1373

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811. http://dx.doi.org/10.1037/0022-3514.69.5.797

Steele, J., James, J. B., & Barnett, R. (2002). Learning in a man's world: Examining the perceptions of undergraduate women in male-dominated academic areas. *Psychology of Women Quarterly, 26,* 46–50.

Stone, J. (2002). Battling doubt by avoiding practice: The effects of stereotype threat on self-handicapping in white athletes. *Personality and Social Psychology Bulletin, 28,* 1667–1678. http://dx.doi.org/10.1177/014616702237648

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 6313–6317. http://dx.doi.org/10.1073/pnas.1221764110

Thompson, J., Morton, J., & Fraser, L. (1997). Memories for the Marchioness. *Memory, 5,* 615–638. http://dx.doi.org/10.1080/741941482

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Chicago, IL: Nelson Hall.

Walton, G. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science, 23,* 73–82.

Wetzel, C. D. (1975). Effect of orienting tasks and cue timing on the free recall of remember- and forget-cued words. *Journal of Experimental Psychology: Human Learning and Memory, 1,* 556–566. http://dx.doi.org/10.1037/0278-7393.1.5.556

Wigfield, A., & Karpathian, M. (1991). Who am I and what can I do? Children's self-concepts and motivation in achievement situations. *Educational Psychologist, 26,* 233–261. http://dx.doi.org/10.1207/s15326985ep2603&4_3

Woodward, A. E., Jr., & Bjork, R. A. (1971). Forgetting and remembering in free recall: Intentional and unintentional. *Journal of Experimental Psychology, 89,* 109–116. http://dx.doi.org/10.1037/h0031188

Received March 26, 2016
Revision received October 29, 2016
Accepted November 10, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 6, 826–841

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000177

Peer Victimization Trajectories From Kindergarten Through High School: Differential Pathways for Children's School Engagement and Achievement?

Gary W. Ladd, Idean Ettekal, and Becky Kochenderfer-Ladd
Arizona State University

This investigation's aims were to map prevalence, normative trends, and patterns of continuity or change in school-based peer victimization throughout formal schooling (i.e., Grades K–12), and determine whether specific victimization patterns (i.e., differential trajectories) were associated with children's academic performance. A sample of 383 children (193 girls) was followed from kindergarten (M_{age} = 5.50) through Grade 12 (M_{age} = 17.89), and measures of peer victimization, school engagement, academic self-perceptions, and achievement were repeatedly administered across this epoch. Although it was the norm for victimization prevalence and frequency to decline across formal schooling, 5 trajectory subtypes were identified, capturing differences in victimization frequency and continuity (i.e., high-chronic, moderate-emerging, early victims, low victims, and nonvictims). Consistent with a chronic stress hypothesis, high-chronic victimization consistently was related to lower—and often prolonged—disparities in school engagement, academic self-perceptions, and academic achievement. For other victimization subtypes, movement into victimization (i.e., moderate-emerging) was associated with lower or declining scores on academic indicators, and movement out of victimization (i.e., early victims) with higher or increasing scores on these indicators (i.e., "recovery"). Findings provide a more complete account of the overall prevalence, stability, and developmental course of school-based peer victimization than has been reported to date.

Keywords: peer victimization, trajectories of peer victimization, peer relations, school engagement, achievement

Introduction

Bullying and peer victimization in educational settings has become a national public concern in part because youth who are victimized by schoolmates—particularly across multiple school years (Troop-Gordon & Ladd, 2005)—evidence a plethora of psycho-social and scholastic adjustment problems (see Ettekal, Kochenderfer-Ladd, & Ladd, 2015; Nakamoto & Schwartz, 2010). Peer victimization has been defined as being bullied or aggressed upon repeatedly and over time by one or more students (Juvonen & Graham, 2014; Olweus, 1999), and has been operationalized by assessing how frequently youth are the recipients of peers' aggressive acts (e.g., see Ladd & Kochenderfer-Ladd, 2002).

The link between peer victimization and academic performance has been examined less thoroughly than its association with other aspects of development, such as child health and psychological adjustment (see Ettekal et al., 2015). Even though research on peer victimization in school contexts is ongoing and has been a source of important discoveries (see Juvonen & Graham, 2014), insight into this phenomenon and its links with children's academic development could be enhanced if investigative attention were focused on three pivotal objectives. First, a more complete descriptive account is needed of the prevalence, stability, and developmental course of peer victimization across the entire period of formal schooling. At present, and as detailed more completely below, more is known about some intervals of schooling (e.g., Grades K–5; 7–8) than others (e.g., Grades 9–12). A broader, more accurate picture of these dimensions could therefore be obtained by mapping both the prevalence and normative (mean) levels of peer victimization in Grades K–12. Second, more needs to be known about intraindividual differences in the continuity or discontinuity of peer victimization across all of the formal school years. We know that some children are victimized earlier versus later in their schools careers, and for longer versus shorter periods of time (reviewed below), but no attempt has been made to determine whether there are specific classes of children who exhibit stable versus increasing or decreasing victimization trajectories from Grades K through 12. Third, along with differences in

This article was published Online First January 30, 2017.

Gary W. Ladd, T. Denny Sanford School of Social and Family Dynamics and Department of Psychology, Arizona State University; Idean Ettekal and Becky Kochenderfer-Ladd, T. Denny Sanford School of Social and Family Dynamics, Arizona State University.

This investigation was conducted as part of the Pathways Project, a larger longitudinal investigation of children's social, psychological, and scholastic adjustment in school contexts that is supported by the National Institutes of Health (1 RO1MH-49223, 2-RO1MH-49223, R01HD-045906 to Gary W. Ladd). Special appreciation is expressed to all the children and parents who made this study possible, and to members of the Pathways Project for assistance with data collection.

Correspondence concerning this article should be addressed to Gary W. Ladd, Department of Psychology, T. Denny Sanford School of Social and Family Dynamics, Arizona State University, P.O. Box 873701, Tempe, AZ 85287-2502. E-mail: Gary.Ladd@asu.edu

the temporal patterning of victimization, there is a need to understand how differences in victimization trajectories are associated with children's academic performance across the formal school years.

Accordingly, this investigation's specific aims were to (a) map prevalence and normative trends in peer victimization across the entire period of formal schooling, (b) identify patterns of continuity or discontinuity (i.e., distinct victimization trajectory classes) across Grades K through 12, and (c) determine the extent to which specific trajectory classes (i.e., differential victimization trajectories) are associated with aspects of children's academic performance over the course of their entire school careers. The latter two aims were examined in the context of factors (e.g., gender, SES, race, school transition timing) that, logically, might be related to the identification of trajectory classes and their associations with children's academic adjustment.

Prevalence and Normative Trends

Prevalence denotes the proportion of persons in a given population that possess a particular characteristic or malady. Although prevalence estimates for peer victimization vary, and some periods of schooling have been investigated more than others, evidence suggests that the occurrence of peer victimization varies by age or grade level. For example, extant findings imply that victimization tends to be more widespread in earlier as opposed to later years of schooling (Nylund, Bellmore, Nishina, & Graham, 2007; Reavis, Keane, & Calkins, 2010; Rudolph, Troop-Gordon, Hessel, & Schmidt, 2011). These findings have led some to conclude that victimization likely "peaks" during middle school (e.g., Nylund et al., 2007). Prevalence also has been reported to vary by gender, but evidence of this difference often has been mixed and substantiation weak. Some have found that more boys than girls are victimized (e.g., Snyder et al., 2003; Sullivan et al., 2006; Wolke et al., 2001), particularly when bullying is perpetrated directly or physically rather than indirectly or relationally (e.g., Delfabbro et al., 2006; Paquette & Underwood, 1999; Storch et al., 2003). Others have found that more girls than boys are victimized (e.g., Seals & Young, 2003). In general, gender differences in prevalence tend not to be statistically robust.

Differential Victimization Trajectories

The question of whether there are distinct classes or subtypes of children who exhibit stable, increasing, or decreasing patterns of victimization across time has been investigated, but not across the entire period of formal schooling. Those who have addressed this question have done so using concurrent, cross-sectional, and short-term longitudinal designs—that is, by gathering data at the beginning and end of multiyear intervals (e.g., 5th and 8th grades; Scholte, Engles, Overbeek, de Kemp, & Haselager, 2007), or at regular intervals (repeatedly) across two to four contiguous grades (e.g., K–3rd grades, Kochenderfer-Ladd & Wardrop, 2001; 2nd–5th grades, Rudolph et al., 2011; 3rd–5th grades, Biggs et al., 2010; 5th–7th grades, Goldbaum, Craig, Pepler, & Connolly, 2003). Accordingly, the current state of knowledge about classes of children who traverse differing (i.e., stable, increasing, or decreasing) victimization trajectories across grades is limited in the sense that more is known about some intervals of schooling (e.g., Grades K–5; 7–8) than others (e.g., Grades 9–12).

In fact, when specific periods of schooling have been examined, the types of trajectory classes identified differ. Across mid to late grade school (i.e., Grades 3–6), investigators have identified between 3 and 5 trajectory subtypes. Boivin, Petitclerc, Feng, and Barker (2010) followed children from Grade 3 through 6 and identified three distinct trajectories: rarely victimized (low-stable; 85.5% of sample; 49% boys), victimized and increasing (moderate-increasing; 10% of sample; 69% boys), and highly victimized and declining (high-decreasing; 4.5% of sample; 54.5% boys). Boys significantly outnumbered girls only in the moderate-increasing group. Biggs et al. (2010), in contrast, tracked children across nearly the same grade levels (Grades 3–5) and found five victimization trajectory groups: low (56.2%), moderate (31.7%), decreasing (5.9%), increasing (4.0%), and chronic (2.1%). Gender distributions were not reported, but gender by subtype analyses revealed no significant sex differences.

With adolescents, followed from late grade school into the middle school, investigators have identified 3 to 4 trajectory subtypes. After gathering data across Grades 5 through 7, Goldbaum et al. (2003) identified four trajectory groups that they labeled nonvictims (low; 88%; 50% males), desisters (high-decreasing; 6%; 58% males), late onset (low-increasing; 5%; 59% males), and stable victims (consistently high; 1%; 60% males). The proportion of males versus females within each group did not differ statistically. In comparison, using data collected online rather than in schools, Sumter, Baumgartner, Valkenburg, and Peter (2012) identified three victimization subtypes with adolescents who were followed from ages 12 to 17. The identified trajectory classes were labeled low (48%), moderate (45%), and high (6%), and all subtypes evidenced declines in victimization over time. Boys and girls were "almost equally represented" in the low and moderate subtypes, but fewer boys (39%) were in the high group.

Thus, at present, research on trajectory classes paints an inconsistent picture of developmental trends. Comparisons across studies are made difficult because trajectory classes have been mapped using different data sources (e.g., self vs. peer reports) and different schooling intervals (i.e., periods of schooling, grades within periods). However, one empirically supported inference is that whereas some youth are seldom abused by peers, others are chronically victimized, and still others experience discontinuities such as progressively increasing or decreasing peer victimization. Other seemingly defensible inferences are that (a) there is a greater propensity for peer victimization to decline (both normatively and by subtypes) rather than increase as youth progress through school, and (b) when increasing trajectory classes are identified, these subtypes occur during earlier rather than later periods of schooling.

Significant gender differences in trajectory class membership appear rare, and when reported, do not follow a consistent pattern. With grade-schoolers, Boivin et al. (2010) found more boys than girls in a moderate-increasing trajectory class, but Biggs et al. (2010) found no gender differences for groups characterized by moderate or increasing victimization. With adolescents, Sumter et al. (2012), found more girls than boys in a high victimization

subtype, but Goldbaum et al. (2003) found no such difference for highly victimized youth.

Associations Among Victimization Trajectories and Academic Performance

When victimization's association with academics has been investigated, achievement has been targeted more often than other indicators of educational performance or adjustment. A meta-analysis of existing evidence (e.g., Nakamoto & Schwartz, 2010) suggests that there is a small but significant negative association between victimization and children's achievement.

The absence of a strong relation between victimization and achievement is perhaps understandable because the determinants (e.g., intelligence, parent education, SES) of scholastic attainment (e.g., summative indictors such as grades, achievement test scores, etc.) are diverse, and some may be more influential than peer maltreatment. Furthermore, when viewed from the perspective that victimization precedes scholastic difficulties (currently, the most corroborated model; see Nakamoto & Schwartz, 2010; Schwartz, Gorman, Nakamoto, & Toblin, 2005), it can be argued that achievement may not be the most proximal or sensitive indicator of victimization's academic consequences. Rather, victimization may be more closely linked with other formative aspects of children's educational experience, including the feelings, motivations, and behaviors they develop toward school (e.g., facets of school engagement), and the perceptions they develop of their academic abilities (i.e., perceived academic competence).

Moreover, it is conceivable that exposure to victimization in school may be particularly disruptive during the foundational periods, such as the grade school years, when children are first formulating and eventually solidifying their school-related attitudes, motivations, perceptions, and behaviors (see Ladd & Dinella, 2009; Ladd, Buhs, & Seid, 2000; Ladd, Herald-Brown, & Reiser, 2008). If so, then it would be expected that victimization, particularly when chronic, would evidence stronger relations with school engagement during earlier rather than later school years.

School engagement. Three forms of school engagement have been linked with learning and achievement: cognitive, behavioral, and emotional (see Fredricks, Blumenfeld, & Paris, 2004; Ladd, Herald-Brown, & Kochel, 2009). Whereas cognitive engagement has been construed as students' level of processing or intellectual effort during learning tasks, behavioral engagement refers to actions such as taking initiative, participating cooperatively, manifesting effort and persistence, adopting classroom norms, and staying out of trouble (Birch & Ladd, 1997; Buhs & Ladd, 2001; Finn, 1989). Emotional engagement has been defined as students' attitudes or sentiments toward school and has been operationalized in terms of children's feelings about peers, teachers, schoolwork, or their overall affective reactions to school (Ladd et al., 2000; Skinner, Wellborn, & Connell, 1990).

Although arguably important, the association between peer victimization and school engagement has been understudied. This is unfortunate because a substantial case can be made for victimization as a determinant of school disengagement. In particular, the experiences engendered in victimization (e.g., punishing interactions, physical harm, embarrassment) may decrease emotional engagement by causing children to develop negative school-related attitudes (i.e., dislike of school) and motivations (i.e., desiring or seeking to avoid school). Support for this hypothesis includes evidence indicating that young victimized children, as compared to their nonvictimized counterparts, display lower school liking and higher school avoidance (e.g., Buhs, Ladd, & Herald, 2006; Kochenderfer & Ladd, 1996). Unfortunately, evidence that speaks to these relations largely is limited to the early school years.

Peer victimization also may undermine children's behavioral engagement in classrooms. In particular, this type of maltreatment may discourage independent participation, or children's propensity to initiate and actively participate in classroom activities. Independent participation has been hypothesized to be a contributor to classroom learning and achievement (see Finn, 1989, 1993; Fredricks et al., 2004; Ladd et al., 2000), and evidence corroborates this premise (see Buhs & Ladd, 2001; Buhs et al., 2006; Finn, 1989; Ladd, Birch, & Buhs, 1999).

Academic self-perceptions. Peer victimization also may color children's perceptions of their academic competence. Compared to nonvictimized children, those who are victimized may be less likely to receive peer support, including classmates' affirmation for their academic skills and accomplishments. These children, moreover, may often receive negative messages from peers about their academic competence or worth as learning companions (e.g., teased about schoolwork; disparaged during peer-mediated learning activities; etc.). Although this hypothesis has not been well investigated, the available evidence is consistent with expectation. In a concurrent study conducted with sixth graders, Thijs and Verkuyten (2008) found that peer victimization and perceived academic competence were negatively correlated. More remains to be learned about victimization's association with children's sense of their academic competence.

Overview of Investigative Aims and Hypotheses

To address this study's three specific aims, participants were followed from grades kindergarten through Grade 12. Repeated assessments were made of children's peer victimization, school engagement, academic self-perceptions, and achievement.

Aim 1: Profile prevalence and normative trends in peer victimization across Grades K–12. Previously reported prevalence trends, even though estimated across limited periods of schooling, led us to hypothesize that declines would be evident in peer victimization levels—both sample-wide and by gender—across the K to 12 school years. Although normative declines in victimization were expected for both genders, past findings led us to expect that—if reliable sex differences were detected—boys would have higher levels of peer victimization than girls (e.g., there would be gender differences in the intercepts but not the slopes of children's normative victimization trajectories). Because no one has profiled normative trends across the entirety of formal schooling, it was specifically of interest to pinpoint the time(s) at which, during the course of children's school careers, normative fluctuations occur in peer victimization.

Aim 2: Examine intraindividual differences in temporal patterns and identify distinct victimization trajectory classes across Grades K through 12. Although investigators have yet to map victimization trajectories across more than a few grade levels, corroboration among previously documented patterns led us to hypothesize that at least three types of trajectory classes would

be identified, including *nonvictims* (i.e., low-stable; Biggs et al., 2010; Boivin et al., 2010; Goldbaum et al., 2003; Sumter et al., 2012) and *high-chronic* victims (or high-decreasing; see Boivin et al., 2010; Goldbaum et al., 2003; Sumter et al., 2012). A high-decreasing rather than a high-stable trajectory was anticipated because this pattern was documented by investigators who utilized the longest follow-through designs (Boivin et al., 2010; Sumter et al., 2012), and, normatively, it was expected that victimization levels would decline across grades. A *moderate* victimization trajectory was also expected because some form of this pattern has been documented in all of the previous longitudinal investigations. However, no hypotheses were made about the direction of this trajectory because findings for this subtype have been inconsistent across studies (i.e., variously reported as stable, increasing, and decreasing). Whether or not an *increasing* victimization trajectory would be identified was treated as an empirical question because this pattern has emerged only sporadically, and in different forms (e.g., low increasing; moderate increasing). Gender differences in trajectory group membership were not anticipated, given that such findings have been rare or inconsistent.

Aim 3: Probe the links among victimization trajectories and academic performance across Grades K–12. Expectations about these linkages were based on the above-articulated rationales and the overarching hypothesis that victimization impedes children's academic engagement, perceptions, and performance. When examined by trajectory classes, academic differences were expected to conform to a chronic stress hypothesis—that is, be largest (most discrepant) between children in trajectories indicative of higher (i.e., magnitude) and longer (i.e., chronicity) victimization patterns (e.g., high-chronic/decreasing subtype) than those who exhibited static, nonoccurring patterns of victimization (i.e., nonvictims, or low-stable subtype). Accordingly, when contrasted with nonvictims, academic engagement, perceptions, and achievement scores were expected to be largely discrepant (significantly lower) for the high-chronic subtype, and moderately discrepant (but still significantly lower) for moderate subtypes.

In the event that declining victimization trajectory subtypes were identified, two potential effect patterns were envisioned. The first was consistent with a "recovery" hypothesis (see Kochenderfer & Ladd, 1996), which postulates that, as victimization diminishes (i.e., children approach nonvictim status), so do the processes (stressors) that inhibit academic performance, which in turn enables children to academically reengage, develop more positive views of their academic competence, and achieve in school. Recovery of this type, however, would not be anticipated for subtypes where declines were not large enough to eradicate children's victimization experiences (e.g., eliminate exposure to stress processes). The second was based on a "scar" hypothesis (see Kochel, Ladd, & Rudolph, 2012; Rohde, Lewinsohn, & Seeley, 1990), which implies that victimization's stress- and coping-related effects endure beyond its cessation and, thus, continue to impede children's school engagement and achievement.

If increasing victimization trajectory classes were identified, it was expected that such subtypes would evidence significantly lower or decreasing patterns of school engagement, perceived academic competence, and achievement. Logically, increasing victimization trajectories signify the exacerbation of maltreatment and its effects (e.g., stress).

Method

Participants

Participants were 383 children (193 girls and 190 boys) who were recruited into a longitudinal study as they entered kindergarten (M_{age} = 5.50) and followed yearly until Grade 12 (M_{age} = 17.89). IRB approval was obtained at the study's inception and renewed in all subsequent years. School district consent was obtained prior to recruitment, and 95% of parents provided written informed consent for their child's participation. Approximately 81% of children made the transition to middle or junior high school, and 19% remained in the same school from K to Grade 8. Approximately 77.8% of children were Caucasian, 17.8% African American, and 4.4% Hispanic, biracial, and other backgrounds. About 24.5% came from families with low annual incomes ($0–$20,000), 36.8% had low to middle incomes ($20,001–$50,000), and 38.7% had middle to high incomes (over $50,001).

Procedure

A repeated-measures, multi-informant design was utilized, and all measures were administered in the spring of the school year. From Grades K to 12, children provided self-report data about peer victimization and school engagement (i.e., school liking, avoidance) and, beginning in Grade 4, reported about their perceived academic competence. Trained project examiners administered measures in counterbalanced order individually (Grades K–5) or in groups using self-paced questionnaire booklets (Grades 6–12). Examiners provided instructions about how to complete each measure, and encouraged children to report about contemporaneous events, circumstances, perceptions, and so forth Examiners also individually administered standardized reading and math tests from Grade 2 to Grade 12. From Grades K to 12, teachers (K–5: the classroom teacher; 6–12: a subject-area teacher) rated each child's classroom engagement.

The analyses performed in this study utilized peer victimization data that were collected on a yearly basis and academic adjustment indicators that were assessed every other year. Although many of the academic indicators were assessed yearly, specifying growth models with 13 waves of data led to estimation problems in some of the models. Thus, to reduce model complexity and maintain consistency across criteria, scores from every other year (e.g., K and Grades 2, 4, 6, 8, 10 and 12, when available) were used for the academic indicators. Demographic information was collected from parents at the outset of the study.

Measures

Child reports of peer victimization were obtained using a previously validated measure. Multiple facets of children's school-related engagement, self-perceptions, and achievement were assessed using established child- and teacher-report instruments. For all study variables, descriptive (range, means, *SD*s,) and reliability (alphas) statistics are reported in Table 1.

Table 1
Descriptive Statistics (Range, Observed Means, and Standard Deviations) and Scale Reliabilities for Peer Victimization and Academic Adjustment Variables (N = 383)

Grade	*n*	Min	Max	*M*	*SD*	α
		Peer victimization				
K	382	1.00	5.00	2.24	1.15	.75
1	383	1.00	5.00	2.26	1.07	.73
2	382	1.00	5.00	2.17	1.04	.77
3	381	1.00	5.00	2.08	1.00	.74
4	370	1.00	5.00	1.95	.91	.80
5	371	1.00	5.00	1.89	.84	.80
6	364	1.00	4.50	1.99	.82	.70
7	362	1.00	5.00	1.59	.82	.86
8	349	1.00	5.00	1.45	.71	.86
9	306	1.00	4.00	1.38	.61	.83
10	287	1.00	4.50	1.37	.62	.82
11	296	1.00	3.75	1.26	.50	.79
12	276	1.00	5.00	1.31	.56	.80
		School liking				
K	383	1.00	5.00	4.17	1.14	.87
2	383	1.00	5.00	4.19	.98	.85
4	372	1.00	5.00	3.37	1.07	.87
6	368	1.00	5.00	3.31	.97	.88
8	349	1.00	5.00	3.26	1.04	.89
10	287	1.00	5.00	3.12	.99	.90
12	276	1.00	5.00	3.14	.98	.91
		School avoidance				
K	383	1.00	5.00	2.77	1.44	.71
2	383	1.00	5.00	2.42	1.31	.76
4	372	1.00	5.00	2.60	1.15	.77
6	368	1.00	5.00	2.70	1.08	.77
8	349	1.00	5.00	2.49	1.09	.81
10	287	1.00	5.00	2.49	1.05	.80
12	276	1.00	5.00	2.48	1.10	.81

Grade	*n*	Min	Max	*M*	*SD*	α
		Academic competence				
4	372	1.00	4.00	3.00	.77	.70
6	364	1.00	4.00	3.16	.68	.73
8	350	1.00	4.00	3.20	.68	.81
10	287	1.00	4.00	3.06	.68	.83
12	276	1.00	4.00	3.12	.67	.83
		Independent participation				
K	383	1.00	3.00	2.37	.63	.89
2	382	1.00	3.00	2.33	.64	.89
4	361	1.00	3.00	2.38	.50	.73
6	346	1.00	3.00	2.34	.61	.86
8	329	1.00	3.00	2.25	.65	.88
10	233	1.00	3.00	2.29	.61	.84
12	225	1.00	3.00	2.32	.54	.82
		Math performance				
2	383	46.00	155.00	96.26	12.61	—
4	370	48.00	144.00	100.22	12.74	—
6	369	56.00	153.00	100.22	16.33	—
8	335	63.00	137.00	101.81	14.76	—
10	282	60.00	135.00	99.02	14.58	—
12	274	61.00	130.00	97.92	13.91	—
		Reading performance				
2	382	47.00	140.00	103.31	14.33	—
4	371	58.00	144.00	103.35	14.06	—
6	370	57.00	144.00	103.03	14.72	—
8	336	57.00	139.00	104.57	14.42	—
10	273	64.00	131.00	104.99	12.61	—
12	268	63.00	128.00	106.72	13.93	—

Peer victimization. Self-reports of victimization best suited this investigation's aims and longitudinal design. In contrast to peer or teacher reports, self-reports have the advantage of providing (a) frequency rather than consensus data, (b) greater rater consistency across grades, (c) scores that are more sensitive to changes in victimization and less affected by reputational biases, and (d) data that reflect children's experiences across multiple school contexts (e.g., classroom, bus, lunchrooms, schoolyard; see Furlong et al., 2010; Olweus, 2010).

Children completed a four-item peer victimization scale (Kochenderfer & Ladd, 1996) to assess the frequency (1 = *almost never*, 2 = *a little*, 3 = *sometimes*, 4 = *a lot*, and 5 = *almost always*) with which they had experienced physical ("Does anyone in your class ever hit you at school?"), verbal ("Does anyone in your class ever say mean things to you at school?"), relational ("Does anyone in your class ever say bad things about you to other kids at school?"), and general victimization ("Does anyone in your class ever pick on you at school?"). Scores were calculated by averaging ratings across the four items at each grade level (αs = .73 to .86).

Confirmatory factor analysis (CFA) was used to assess measurement invariance. Longitudinal measurement models were specified such that peer victimization indicators at each of the 13 waves (K–12) were used as the observed indicators and loaded onto latent factors representing peer victimization over time. First, configural invariance was assessed by specifying a baseline model in which factor loadings and intercepts were unconstrained. This model had adequate model fit (χ^2 = 1665.53, *df* = 1130, $p < .001$; RMSEA = .04; SRMR = .05; CFI = .93). Second, weak measurement invariance was assessed by specifying a model in which factor loadings for similar items were constrained to be equal over time. This model had adequate model fit (χ^2 = 1780.52, *df* = 1166, $p < .001$; RMSEA = .04; SRMR = .06; CFI = .92). Third, strong factorial invariance was assessed by specifying a model in which the factor loadings and intercepts of similar items were constrained to be equal over time. This model appeared to have inadequate model fit based on several fit indices (χ^2 = 2282.57, *df* = 1214, $p < .001$; RMSEA = .05; SRMR = .14; CFI = .86).

Nested model comparisons were used to contrast these models. Because difference tests based on χ^2 are sensitive to sample size, contrasts were conducted for multiple fit indices (see Cheung & Rensvold, 2002). When testing for weak measurement invariance, although the nested model comparisons based on the χ^2 were statistically significant ($\Delta\chi^2$ = 114.99, *df* = 36, $p < .001$), differences among the other fit indices were trivial ($\Delta RMSEA$ < .01, $\Delta SRMR$ < .01, ΔCFI = −.01). Cheung and Rensvold (2002) recommend that a difference score of −.01 or less on the CFI can be used as a cutoff for retaining the more parsimonious model.

Thus, it appeared that weak measurement invariance was a reasonable assumption, indicating that the nature of peer victimization remained stable over time. Nested model comparisons indicated that imposing strong measurement invariance resulted in

a significant decline in model fit ($\Delta\chi^2 = 502.05$, $df = 48$, $p < .001$, $\Delta RMSEA = .01$, $\Delta SRMR = .08$, $\Delta CFI = -.06$). Within a longitudinal measurement invariance test, the lack of strong invariance indicated mean-level changes in the peer victimization indicators over time. Although this invariance test does not indicate exactly how mean-level changes were occurring over time, by using latent growth modeling to assess normative trends in peer victimization from K-12, it was possible to more concisely ascertain the nature of these mean-level changes.

Prevalence was calculated as the proportion of children sampled at each grade, from K to 12, who had victimization scores high enough to be considered "victimized." To examine prevalence by victimization frequency, two classification criteria were created that were referenced against item scaling: moderate (i.e., children with scores between 2.00 and 3.50) and severe (i.e., children with scores above 3.50). Scores of 2.00 were for children who, averaging over the four forms of victimization, indicated that they had experienced victimization "a little." Scores greater than 3.50 were for children who indicated that at least two of the four forms of victimization happened to them "a lot."

School engagement. Assessed aspects of school engagement included emotional (i.e., child-reported school liking/disliking), motivational (i.e., child-reported school avoidance), and behavioral (i.e., teacher reports of classroom independent participation) components. A 7-item version of the School Liking and Avoidance Questionnaire (Ladd, 1990; Ladd & Price, 1987) was used to assess emotional (i.e., school liking; 4 items) and motivational (school avoidance; 3 items) aspects of school engagement. Children rated each item using a 5-point scale (1 = *almost never*, 2 = *a little*, 3 = *sometimes*, 4 = *a lot* and 5 = *almost always*). Example school liking items included "Are you happy at school?" and "Do you like being in school?" School avoidance items included "Do you wish you did not have to go to school?" and "Do you ask your parents to let you stay home from school?"

CFA was used to determine whether items from the school liking and avoidance subscales assessed distinct dimensions of school engagement (i.e., school liking and avoidance). For each of seven data waves (Grades K, 2, 4, 6, 8, 10 and 12), a 1-factor model (combining all school liking and avoidance items) was compared to a 2-factor model (school liking and avoidance items specified as correlated factors). For all waves, the 2-factor model had adequate model fit (*RMSEA*s < .08; *SRMR*s < .04; *CFI*s > .98) and fit the data better than the 1-factor model (*RMSEA*s = .13–.22; *SRMR*s = .05–.10; *CFI*s = .83–.93). Accordingly, in subsequent analyses, school liking and school avoidance were examined as distinct constructs. Subscale scores were created by averaging ratings across component items.

The Independent Participation (4 items) subscale of the Teacher Rating Scale of School Adjustment (see Birch & Ladd, 1997; Ladd, Kochenderfer, & Coleman, 1996) was used to assess children's behavioral classroom engagement. Example items include "shows initiative" and "works independently," and teachers rated each item on a 3-point scale (1 = *doesn't apply*, 2 = *applies sometimes*, and 3 = *certainly applies*). Seven waves of participation data were utilized (Grades K, 2, 4, 6, 8, 10 and 12), and scores for this measure exhibited adequate reliability.

Perceived academic competence. Children's self-perceptions in the academic domain were evaluated with a 4-item child-report version of the widely used and validated Perceived Academic Competence subscale of the Perceived Competence Scale for Children (Harter, 1982). Each item is presented using a structured alternative response format (e.g., "Some kids feel like they are very good at their schoolwork . . . but other kids worry about whether they can do the schoolwork assigned to them."). Children are instructed to choose the alternative that is more like them, and then rate whether that response is "*sort of true*" or "*really true*" for them. Item responses are scored on a 4-point scale (1–4) and then averaged such that higher scores denote greater perceived academic competence. Five waves of academic competence data were used in this study (Grades 4, 6, 8, 10, and 12) and, for each wave, scale reliability was adequate.

Academic achievement. Reading and math achievement were assessed using corresponding subscales of the Wide Range Achievement Test (WRAT; Wilkinson, 1993). WRAT subscale items are scored as incorrect or correct (scores = 0, 1). The WRAT is suitable for children in Grades 2–12, has adequate psychometric properties, and has been normed and validated on national samples (Hughes, 1987; Wilkinson, 1993). For each wave, standard scores for reading and math were computed for each participant using the scale developer's scoring procedures. WRAT scores from Grades 2, 4, 6, 8, 10 and 12 were utilized.

Data Analysis Plan

First, multiple-group latent growth modeling (Mplus; Muthén & Muthén, 1998–2010) was performed to assess normative trends in peer victimization from kindergarten to Grade 12 and to determine whether these trends varied by gender. Second, growth mixture modeling (GMM) was performed to identify classes of children with similar victimization trajectories from K to 12. Third, study hypotheses were evaluated by examining time-varying differences in academic performance for children in different victimization class trajectories, with the nonvictim subtype serving as a referent group. For each academic indicator, growth models were used to contrast patterns of continuity or change among victim trajectory classes, and premises about potential effect patterns (i.e., temporal differences consistent with recovery, scar, or developmental hypotheses) were evaluated by comparing linear growth models with piecewise growth models (via nested model comparisons). To reduce the complexity of the latent growth models, and keep the models consistent across criteria, scores from every other year (e.g., K and Grades 2, 4, 6, 8, 10 and 12, when available) were used for the academic indicators.

Results

Missing Data Analyses

Examination of missing data and participant attrition revealed that, for all study variables, 12.1% of the data were missing. Attrition increased with time (0% in Grade 2, 1.0% in Grade 4, 0.8% in Grade 6, 3.7% in Grade 8, 7.8% in Grade 10, and 9.9% in Grade 12). In all, by Grade 12, 23.2% of participants had dropped out. A series of univariate *t* tests were performed to examine the associations between children's gender, race, and socioeconomic status (e.g., family income) and the likelihood of having either missing data on a specific measure or dropping out of the study. Results showed that boys were more likely than girls to drop out;

however, there were no differences by race or family income. Moreover, boys, African Americans, and children with lower family incomes were more likely to have missing data on self-report measures, but not teacher reports, in Grades 10 and 12, but these differences were small in magnitude.

Missing data were handled in Mplus using full information maximum likelihood (FIML) estimation. This approach is advantageous compared to more traditional missing data techniques because it includes all participants in the analyses (n = 383) regardless of whether they had missing data or dropped out of the study (Enders, 2010). In order for FIML to provide accurate and unbiased parameter estimates, observable causes of missingness should be included within the specified models. Therefore, in addition to examining gender and middle school transition effects, race and SES were included in the growth models. When possible, race and SES were specified as auxiliary variables using the *auxiliary* command in Mplus (which uses a *saturated correlates* model; see Enders, 2010). However, this command cannot be used in mixture models (Muthén & Muthén, 1998–2010), in which case they were specified as covariates in the GMMs. Although the addition of covariates in a GMM may impact the identification of classes (see Muthén, 2004), the peer victimization trajectory classes that were identified were nearly identical when comparing models with and without covariates.

Prevalence of Peer Victimization

The proportions of children sampled at each grade who had scores that fit the criteria for moderate (2.00 to 3.50) and severe (>3.50) victimization are shown in Table 2. Children were categorized as low-victims if their peer victimization scores were less than 2.00. Whether estimated as moderate or severe, prevalence rates for peer victimization declined across grades.

Normative Trends in Peer Victimization by Gender

Normative trends in peer victimization were assessed from Grades K to 12 using a series of multiple-group latent growth models that were designed to assess gender differences in time-varying changes in peer victimization. Fit for all growth models was deemed adequate if RMSEA < .06 and SRMR < .08 (Hu & Bentler, 1999). Because software packages use an inappropriate baseline model to compute the CFI in growth curve models (Wu, West, & Taylor, 2009), this index was not interpreted for these models.

A model-building strategy was used such that a multiple-group *linear* growth model was first specified that included latent intercept and slope factors to capture linear changes in peer victimization from kindergarten to Grade 12. This model did not appear to have adequate model fit (RMSEA = 09; SRMR = .11). To assess possible sources of misfit, residual variances and observed means were examined, and a nonlinear latent growth model was specified that included a quadratic effect to capture nonlinear changes in peer victimization over time. This model did not have adequate fit (RMSEA = 08; SRMR = .10). A cubic growth factor was added to the latent growth model; this model had adequate fit (RMSEA = 06; SRMR = .08) and more accurately estimated the observed peer victimization scores than previous models.

To determine whether there were gender differences in normative trends in peer victimization from kindergarten to Grade 12, nested model tests (i.e., Likelihood Ratio Tests; LRTs) were performed. Toward this end, the unconstrained model in which the latent growth (i.e., intercept, slope, quadratic and cubic) factors were estimated for each group (log likelihood = 5033.61) was compared to a constrained model in which constraints were added to make these factors equal for girls and boys (log likelihood = 5039.65). The LRT indicated that constraining these factors between genders resulted in a statistically significant reduction in model fit ($-2^{*}\Delta LL = 12.08$, $\Delta df = 4$, $p = .02$). Follow-up nested model comparisons indicated that gender differences were primarily attributable to intercept differences between boys and girls (log likelihood = −5034.27; $-2^{*}\Delta LL = 1.31$, $\Delta df = 3$, $p = .73$ with unconstrained model and $-2^{*}\Delta LL = 10.77$, $\Delta df = 1$, $p = .01$ with constrained model). Thus, although boys appeared to have significantly higher levels of peer victimization than girls (i.e., intercept differences; M = 2.17 for girls and M = 2.31 for boys), the patterns of change over time between girls and boys were similar (for girls and boys: $M_{slope} = .25$, $M_{quadratic} = -2.64$, $M_{cubic} = 1.50$; see Figure 1).

Table 2
Victimization Prevalence Rates by Grade

Grade	Low victims	Moderate victims	Severe victims
K	41.1	38.0	20.9
1	39.7	41.5	18.8
2	42.1	42.1	15.7
3	42.0	45.4	12.6
4	56.5	33.8	9.7
5	60.6	33.7	5.7
6	50.5	44.5	4.9
7	75.1	21.3	3.6
8	82.2	14.9	2.9
9	83.7	15.4	1.0
10	86.1	12.2	1.7
11	91.2	7.8	1.0
12	88.8	10.5	.7

Note. Low victims ≤ 2. Moderate victims from 2 to 3.5. Severe victims ≥ 3.5.

Differential Victimization Trajectories

To identify groups of children with heterogeneous peer victimization trajectories from kindergarten to Grade 12, a series of growth mixture models (i.e., 2- thru 6-classes) were specified using the 13 (yearly) waves of peer victimization data. Models included intercept, slope, and quadratic latent growth factors. Several criteria were used to determine the optimal solution, and for each model, multiple fit indices were evaluated in addition to examining whether the trajectory classes appeared substantively and conceptually meaningful (Ram & Grimm, 2009; Tofighi & Enders, 2008). A combination of multiple information criteria (i.e., AIC, BIC, and sample-size adjusted BIC referred to as SABIC), the likelihood ratio test (i.e., Lo-Mendell-Rubin likelihood ratio test; LMR-LRT), and classification accuracy were used to assess each model. Models with smaller AIC, BIC, and SABIC values indicate better solutions. A significant p value on the LMR-LRT indicates that a model with k classes has better fit to the observed data than a model with $k-1$ classes. Classification accuracy was

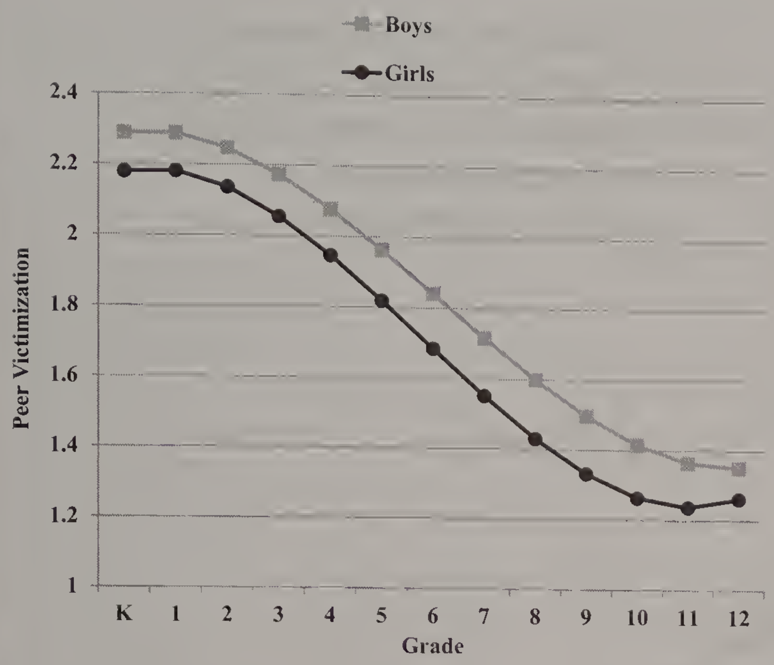


Figure 1. Predicted normative trends in peer victimization from kindergarten to Grade 12 by gender.

assessed by examining the entropy and class assignment probabilities for each model (values closer to 1 indicate more precise classification).

Based on these criteria, the 5-class model appeared to be the optimal solution (see Table 3 and Figure 2). This model had the smallest BIC, second smallest AIC and SABIC, high entropy. and average class assignment probabilities, and the LMR-LRT approached statistical significance (p = .08). Although the 6-class solution had the smallest AIC and SABIC, the addition of a sixth class did not improve model fit compared to the 5-class solution ($2 * \Delta LL = 42.8$, $p = .14$), and the additional class identified in this model was not conceptually distinguishable from the classes identified in the 5-class solution. The 5-class solution (see Figure 2) consisted of 24.0% of children in a high-declining peer victimization trajectory class (labeled *high-chronic victims*), 25.8% in a high steeply declining trajectory class (labeled *early victims*), 17.8% in a moderate increasing victimization trajectory class (labeled *moderate-emerging victims*), 25.8% in a moderate declining victimization class (labeled *low victims*), and 6.5% in a very low victimization trajectory class (labeled *nonvictims*). After determining the optimal model, covariate effects were assessed. High-chronic and moderate-emerging victims were significantly more likely ($ps < .01$) to be boys than nonvictims (65% and 64%, respectively). Making the transition to middle school was not associated with victimization class membership. Nonetheless, in subsequent growth models assessing academic adjustment trajectories, gender and middle-school transition were specified as covariates.

Differential Victimization Trajectories and Academic Adjustment Trajectories

To examine the associations between children's peer victimization trajectories and their academic adjustment trajectories, separate latent growth models were estimated for each of the academic indicators. Peer victimization trajectory class assignments were used to compute a series of dummy coded variables that were then regressed on the latent intercept and slope academic factors. Because both the low and nonvictims trajectory groups appeared to have relatively low levels of peer victimization and the nonvictims group was smaller than other groups, these two groups were combined to increase the size of the referent group in subsequent analyses (referred to as the *low victims* group hereafter).

To explore whether there were developmental differences between the early (K–6) and later grades (6–12), and to provide a more nuanced assessment of the chronic, recovery, and scar hypotheses, for each academic indicator a piecewise growth model was compared to a linear growth model using nested model comparisons. Piecewise models included two latent slope factors and one intercept factor. To test for differences by victimization trajectory classes within the piecewise models, three regression effects were estimated for each class. One effect estimated differences in children's academic trajectories (slopes) from baseline assessment to Grade 6, the second effect estimated slope differences from Grades 6 to 12, and the third effect estimated intercept differences. For each academic adjustment indicator, models were specified twice by adjusting the intercept to estimate differences in children's baseline and Grade 12 academic adjustment. If developmental differences between the early and later grades were pronounced, then it would be expected that there would be significant differences between the two slope factors and between the victimization class regression effects. If these differences were nonsignificant, then a more parsimonious linear growth model was used.

Table 3
Model Fit Indices and Class Proportions for Peer Victimization Trajectories

							Percent of children in each class (and average class assignment probabilities)					
Model	LogL	AIC	BIC	SABIC	Entropy	LMR-LRT	1	2	3	4	5	6
2-Class	−6433.16	12908.32	12991.23	12924.60	.87	937.38**	59.0 (.97)	41.0 (.95)	—	—	—	—
3-Class	−6303.60	12665.20	12779.69	12687.68	.81	251.65**	34.7 (.93)	30.0 (.88)	35.3 (.94)	—	—	—
4-Class	−6209.94	12493.89	12639.97	12522.57	.83	180.06**	34.5 (.92)	24.5 (.87)	34.2 (.91)	6.8 (.96)	—	—
5-Class	−6171.31	12432.63	12610.29	12467.51	.79	72.92	24.0 (.91)	25.8 (.88)	17.8 (.78)	25.8 (.86)	6.5 (.93)	—
6-Class	−6148.38	12402.77	12612.01	12443.85	.78	42.8	23.0 (.87)	27.2 (.87)	13.1 (.78)	23.0 (.83)	7.8 (.85)	6.0 (.94)

Note. LogL = Loglikelihood; AIC = Akaike information criteria; BIC = Bayesian information criteria; SABIC = Sample-size adjusted Bayesian information criteria; LMR-LRT = Lo-Mendell-Rubin likelihood ratio test.
** $p < .01$.

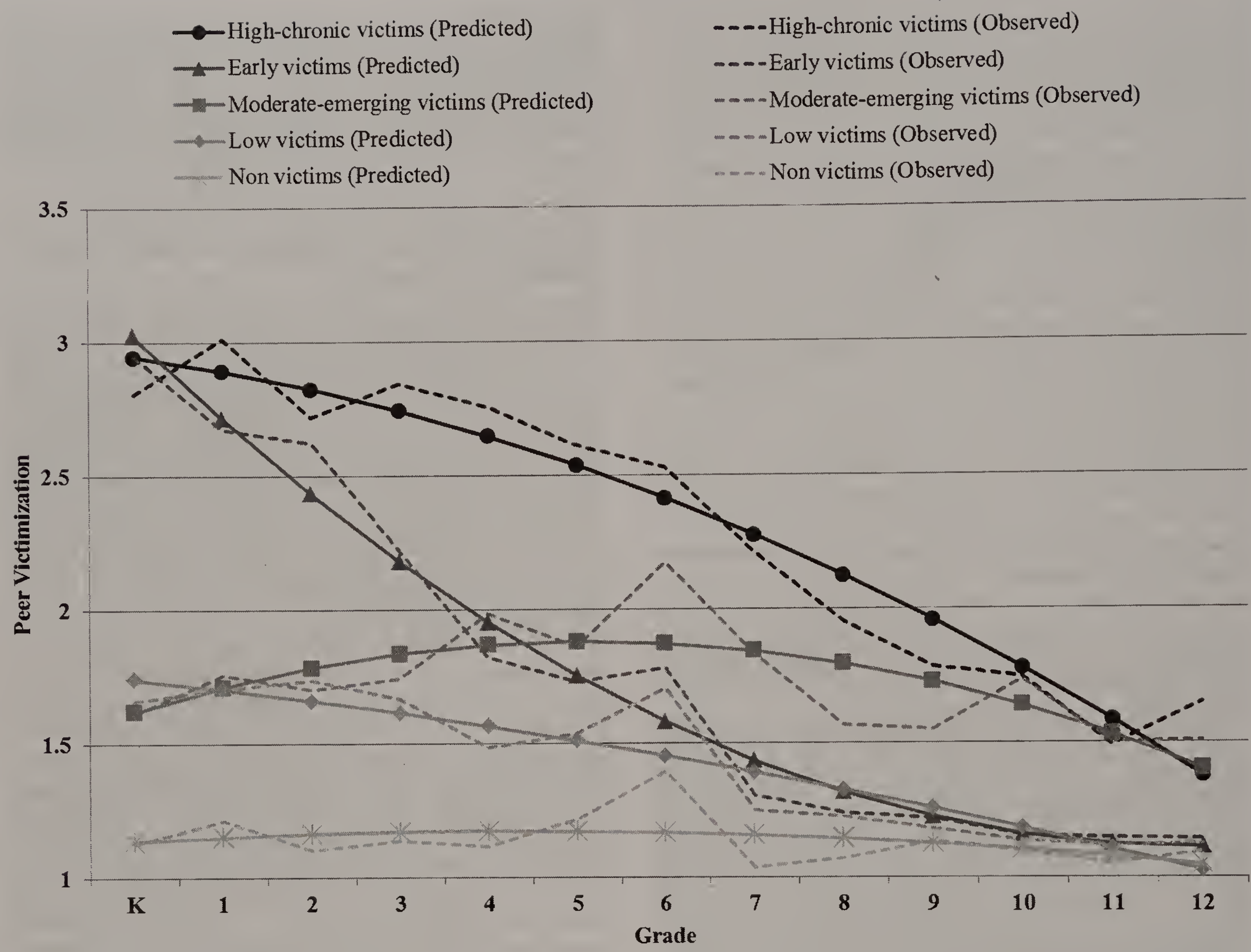


Figure 2. Children's predicted and observed peer victimization trajectories from kindergarten to Grade 12 based on 5-class solution consisting of *high-chronic victims* (24.0%), *early victims* (25.8%), *moderate-emerging victims* (17.8%), *low victims* (25.8%), and *nonvictims* (6.5%).

In addition to examining the effects of children's peer victimization trajectories, these models also accounted for gender differences and whether or not children made the transition to middle school. Gender and middle school transition were specified as covariate (main) effects and regressed on the latent intercept and slope factors. A series of models were also estimated to test for gender by victimization class interaction effects. These interaction effects were consistently nonsignificant and did not improve model fit. Thus, these interaction effects are not reported here, and more parsimonious models are presented without interaction effects. Results (i.e., estimates and significance tests) for these models are presented in Table 4 and illustrated (for interpretative purposes) in Figure 3. Notably, although many of the trajectories illustrated in Figure 3 appeared to be different from one another, suggestive of subtype differences, these differences should be assessed with consideration of the significance tests reported in Table 4 to account for the variability (standard errors) in these estimated trajectories.

School engagement. The conditional piecewise growth model for school liking had adequate fit (χ^2 = 120.89, df = 38, p < .001; RMSEA = .08; SRMR = .05). Furthermore, a nested model comparison ($\Delta\chi^2$ = 120.89, df = 8, p < .001) revealed that the piecewise model had better fit compared to a linear growth model (χ^2 = 196.13, df = 46, p < .001; RMSEA = .09; SRMR = .07). This model revealed a developmental pattern in which children's school liking trajectories were highest in kindergarten and exhibited a steep decline through the grade school years (i.e., Grade 6); however, these trajectories appeared to level off and become more stable in middle and high school (i.e., Grades 6 to 12; see Figure 3 top left). Controlling for gender and the middle school transition, the results indicated that early victims had significantly lower rates of school liking in kindergarten compared to the reference group (i.e., low victims; $M_{intercept}$ = 4.42). However, by Grade 12, high-chronic victims had lower rates of school liking, and the effect for early victims was attenuated. Although there were no significant slope effects (i.e., differences) between victimization classes, the results revealed a decline in school liking for low-victims during the early grade school years (M_{slope} = −.33, p < .001). In addition to the victimization effects, boys had a more significant decline in school liking during the early grade school

Table 4
Estimates for Conditional Growth Models Examining Children's Academic Adjustment Trajectories by Peer Victimization Trajectory Class

Predictors	School liking		School avoidance		Independent participation		Perceived academic competence		Math performance		Reading performance	
	Est.	*SE*	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*
Baseline Effects												
Moderate-emerging	.03	.15	−.09	.19	−.30***	.08	−.16	.10	−2.36	1.79	−5.35**	2.09
Early	−.31*	.13	.63***	.17	−.12	.07	−.08	.09	1.20	1.57	−1.42	1.84
High-chronic	−.24	.14	.26	.17	−.37***	.07	−.49***	.09	−4.54**	1.63	−7.64***	1.91
Gender	.02	.10	.11	.13	−.15**	.06	.12	.07	−2.14	1.23	−1.78	1.44
M.S. transition	−.04	.13	.14	.16	−.08	.07	.08	.08	3.40*	1.57	−2.27	1.80
Grade 12 Effects												
Moderate-emerging	−.20	.17	.12	.18	−.21*	.08	−.32**	.11	−5.80*	2.29	−3.36	2.15
Early	−.17	.14	−.09	.15	−.05	.07	−.07	.09	−1.46	1.98	−.79	1.83
High-chronic	−.32*	.16	.40*	.17	−.35***	.08	−.41***	.10	−7.09***	2.11	−2.91	1.96
Gender	−.22	.11	.11	.12	−.19***	.06	.03	.07	−.20	1.57	.51	1.46
M.S. transition	−.13	.14	.10	.16	−.13	.07	−.03	.09	−4.85*	1.98	−4.68**	1.82
Slope Effects (K-6)												
Moderate-emerging	−.09	.06	.07	.08	.01	.02	−.04	.03	−2.47*	1.10	.15	.79
Early	.06	.06	−.20**	.07	.01	.02	.00	.03	−2.23*	.97	−.23	.69
High-chronic	−.03	.06	−.07	.07	.00	.02	.02	.03	−2.44*	1.01	1.24	.72
Gender	−.10*	.04	.07	.05	−.01	.01	−.02	.02	−.10	.76	.41	.54
M.S. transition	−.02	.06	−.04	.07	−.01	.02	−.03	.03	−2.05*	.96	1.34*	.68
Slope Effects (G6–G12)												
Moderate-emerging	.01	.06	.00	.08	.01	.02	−.04	.03	.50	.58	.56	.61
Early	−.01	.06	−.04	.07	.01	.02	.00	.03	.60	.49	.36	.51
High-chronic	.00	.06	.12	.07	.00	.02	.02	.03	.77	.54	.75	.56
Gender	.02	.04	−.07	.05	−.01	.01	−.02	.02	.71	.40	.49	.41
M.S. transition	−.01	.06	.03	.07	−.01	.02	−.03	.03	−1.39**	.50	−1.70***	.51

Note. For gender, 0 = female, 1 = male; M.S. = middle school; G = grade. For all growth models, the low victim trajectory class served as the reference group.
* $p < .05$. ** $p < .01$. *** $p < .001$.

years than girls and subsequently lower levels of school liking by Grade 12. Making the middle school transition was not associated with students' school liking trajectories.

The conditional piecewise growth model for school avoidance had adequate model fit ($\chi^2 = 84.31$, $df = 38$, $p < .001$; RMSEA = .06; SRMR = .04; see Figure 3 top middle). Furthermore, a nested model comparison ($\Delta\chi^2 = 16.57$, $df = 8$, $p = .03$) revealed that the piecewise model had better fit compared to a linear growth model ($\chi^2 = 100.88$, $df = 46$, $p < .001$; RMSEA = .06; SRMR = .05). Compared to low victims ($M_{intercept} = 2.41$), early victims had significantly higher rates of school avoidance in kindergarten; however, they also had a significant decline in school avoidance during the early schooling years (i.e., K–6) and were not significantly different from low victims by Grade 12. By Grade 12, high-chronic victims had significantly higher school avoidance than low victims.

The conditional piecewise growth model for independent participation resulted in estimation problems, but the linear growth model had adequate model fit ($\chi^2 = 84.12$, $df = 44$, $p < .001$; RMSEA = .05; SRMR = .07; see Figure 3 top right). Compared to low victims ($M_{intercept} = 2.56$), moderate-emerging and high-chronic victims had significantly lower levels of independent participation in kindergarten, which were sustained until Grade 12. In addition to the victimization effects, boys consistently had lower levels of independent participation.

Perceived academic competence. In contrast to the other academic indicators, the baseline assessment for perceived academic competence was not collected until Grade 4. For this reason, a linear rather than a piecewise growth model was used to assess slope differences in children's trajectories from Grades 4 to 12. This model had adequate fit ($\chi^2 = 55.56$, $df = 24$, $p < .001$; RMSEA = .06; SRMR = .06; see Figure 3, bottom left). In Grade 4, compared to low victims ($M_{intercept} = 3.25$), high-chronic victims had lower academic competence, and this effect was sustained until Grade 12. Although significant slope effects were absent, by Grade 12 the moderate-emerging victims also had significantly lower academic competence than low-victims.

Academic achievement. Conditional latent growth models were also estimated to assess children's math and reading performance from Grades 2 to 12. The piecewise model for math performance resulted in adequate fit ($\chi^2 = 81.56$, $df = 27$, $p < .001$; RMSEA = .07; SRMR = .04). Moreover, compared to a linear growth model ($\chi^2 = 213.34$, $df = 35$, $p < .001$; RMSEA = .12; SRMR = .08), the piecewise model exhibited better fit ($\Delta\chi^2 = 131.78$, $df = 8$, $p < .001$). The piecewise model (see Figure 3, bottom middle) revealed a developmental pattern in which children's math performance trajectories increased from Grades 2 through 6, and then declined thereafter. Compared to low victims ($M_{intercept} = 98.0$), high-chronic victims had significantly lower math performance in Grade 2 that was sustained until Grade 12, despite this normative change in scores over time. Moreover, by Grade 12, the moderate-emerging victims also had significantly lower math performance than low victims. Compared to low victims who had a significant increase in math performance from

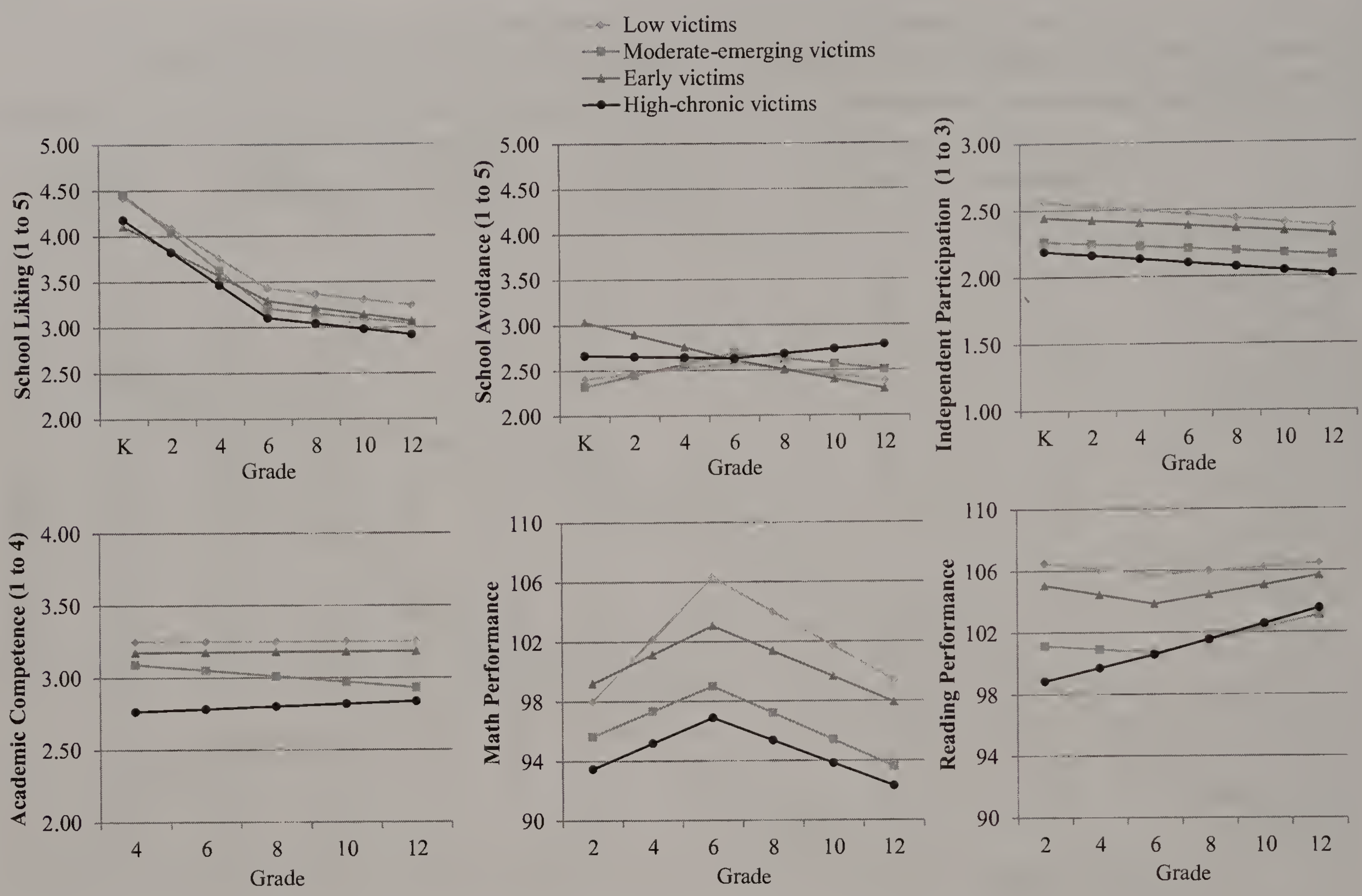


Figure 3. Children's predicted academic trajectories by victimization classes.

Grades 2 to 6 (M_{slope} = 4.16, $p < .001$), all three victims groups (early, moderate-emerging, and high-chronic) were found to have significantly less pronounced gains in math performance during these grades. A middle-school transition effect also was detected, such that children who made the transition showed more significant declines in their performance over time and had lower math performance by Grade 12.

The conditional piecewise growth model for reading performance had adequate model fit (χ^2 = 48.55, df = 27, $p < .01$; RMSEA = .05; SRMR = .02). Moreover, compared to a linear growth model (χ^2 = 70.27, df = 35, $p < .001$; RMSEA = .05; SRMR = .05), the piecewise model exhibited better fit ($\Delta\chi^2$ = 21.72, df = 8, $p < .01$; see Figure 3 bottom right). In Grade 2, compared to low victims ($M_{intercept}$ = 106.50), the moderate-emerging and high-chronic victims had significantly lower reading performance, but these effects were attenuated by Grade 12. Children who made the middle-school transition showed more significant declines in their performance in subsequent years and had lower reading performance by Grade 12.

Discussion

The results of this study make three important contributions to what is known about peer victimization in educational settings. First, the data provide a more complete descriptive account of the overall prevalence, stability, and developmental course of peer victimization across the entire period of formal schooling (i.e., the K–12 school years)—a substantially longer epoch than has been investigated to date. Second, the findings not only corroborate prior research by suggesting that there are multiple subtypes of children who are more or less victimized (i.e., within-sample subtypes), but also extend past evidence by providing a more complete picture of the temporal patterning of victimization (i.e., trajectories throughout formal schooling) for children who are members of specific classes (i.e., subtype trajectories). Third, because multiple indicators of academic performance were examined from Grades K to 12, this study's findings provide novel insights into the long-term associations between specific victimization trajectories, achievement, and focal achievement precursors.

Peer Victimization: Prevalence, Normative Trends, and Differential Trajectories From Grades K to 12

Prevalence and normative trends. Our estimations of victimization's prevalence were consistent with prior investigators' conclusions about age-related or temporal changes (see Nylund et al., 2007; Reavis et al., 2010; Rudolph et al., 2011). First, the proportion of victims identified differed significantly by age or grade level. Second, the prevalence of victimization was higher in earlier as opposed to later school years. Unlike prior findings, however, which have been limited to specific intervals of school-

ing (e.g., Grades K–5; 7–8) and conflicting in their characterization of developmental trends, data from this study suggest that prevalence rates undergo a continuous reduction in magnitude across the entire span of formal schooling. Third, although boys on average appeared to experience somewhat higher levels of victimization than girls, the normative trends for both genders were nearly identical.

At present, there is little basis for understanding why normative estimations of the prevalence of peer victimization show a progressive decline across the school years, as found here. Methodological explanations are conceivable, such as the premise that the observed decline in prevalence is an artifact of self-report assessment (e.g., younger children are more willing to report victimization; children may relax their definitions of peer aggression as they mature, etc.). However, evidence attesting to this instrument's measurement invariance and validity (Ladd & Kochenderfer-Ladd 2002) argue against the credibility of these interpretations.

Alternatively, if prevalence does decline progressively across grades, as our findings and others suggest, then it becomes important to understand its determinants. Unfortunately, as of yet, no theory speaks to developmental vicissitudes in peer victimization. However, in the search for explanations, various maturational and socialization processes deserve consideration, including those that likely decrease children's involvement in bullying and other forms of peer aggression. These include changes in children's maturity (e.g., growth in moral reasoning, perspective taking, empathy, etc.), social environments (e.g., movement toward selective peer environments, peer niche seeking at later grade levels), and socialization processes (e.g., increasing sanctions against bullying and aggression).

Differential trajectories. Not only do our findings suggest that, normatively, the prevalence of peer victimization declines across formal school years, but for the sample as a whole, the results show that frequency—as reflected in average victimization scores—does as well. However, the latter normative trend was not representative of all children's victimization experiences. Rather, substantial intraindividual differences were found in the frequency and continuity of children's peer victimization trajectories across the K–12 school years.

Each of the five victimization trajectories we identified was consistent with expectations or previously reported findings. As hypothesized, two of these subtypes contained participants who were nearly opposites: children who were rarely victimized (i.e., nonvictims), and children who were severely and chronically victimized (i.e., chronic victims; e.g., Biggs et al., 2010; Boivin et al., 2010; Sumter et al., 2012). The third group, termed "early victims" (initially high followed by a steep, continuous decline), had what prior investigators have labeled "rapidly decreasing" or "desister" trajectories (e.g., Biggs et al., 2010; Goldbaum et al., 2003). Of the two moderately victimized subtypes, the low group had a trajectory similar to that reported by Sumter et al. (2012; i.e., "moderate-decreasing"), and the moderate-emerging group followed a course similar to one reported by Goldbaum et al., 2003 (i.e., "late-onset"). Compared to past studies, however, the subtypes identified in this study are significant because, rather than describing trajectories across a few grade levels or isolated schooling epochs, as has been typical, these subtypes characterize victimization experiences that transcend children's entire school careers.

These patterns in the continuity of peer victimization in educational settings present educators with both bad news and good news. First the bad news: Sadly, the discovery of a high-decreasing or chronic victim subtype suggests that, for a substantial number of children (24% of our sample), moderate to severe peer victimization is a stable or enduring part of their educational experience throughout formal schooling. Although the frequency of victimization for children in this subtype declined across grades, as was the norm, it nonetheless remained as high as (and most often higher than) the levels documented for children in all other subtypes. Another piece of bad news was the pattern found for the moderate-emerging victim subtype (17.8% of our sample). These children started school with moderate victimization levels, but their exposure to peer aggression did not decline, as was the norm for all other children, but *increased* across the late elementary and middle school years before diminishing to the level observed for chronic victims. It might be argued that this data pattern lends support to the claim that victimization "peaks" during the transition to middle school (e.g., Nylund et al., 2007). However, the fact that this trajectory was evident in a small proportion of the sample and not characteristic of most children implies that there may only be a subset of children for whom this conclusion applies. For most children, there seems to be a normative decline in victimization throughout this period.

The good news is that we also found groups of children who, although victimized at moderate to high levels as they began school (i.e., early, low subtypes), essentially "recovered" as they moved through the grades. By the time these children reached high school, their average victimization scores were similar to nonvictims. These two groups warrant further investigative attention because understanding what drives desistence could have important implications for prevention and intervention research. Unfortunately, the design of this study precluded opportunities to determine what might account for the desistence exhibited by children in these trajectory subtypes. One hypothesis that merits consideration in future studies is that the children who are represented within these subtypes (i.e., "desisters") possess certain psychological or social resources that allow them to overcome early victimization experiences (e.g., more friendships, social competence, adaptive coping responses, etc.).

Victimization Trajectories and Academic Performance Across Grades K–12

This study's results corroborate the inverse relation between peer victimization and children's academic performance that has been reported previously (see Nakamoto & Schwartz, 2010). Furthermore, the findings extend what is known by clarifying how each of the identified, long-term victimization patterns is related to *specific* aspects of school engagement and achievement.

School engagement. Particular victimization trajectories were found to be associated with all of the investigated aspects of school engagement. School liking was an indicator of children's emotional engagement toward school. Normatively, these feelings became less positive over the course of formal schooling—a trend that is consistent with evidence (see Ladd et al., 2000) suggesting that liking begins high because children initially underestimate the demands of schoolwork, but declines

as they develop more realistic feelings toward school. Differences by gender were found in that boys' school liking declined at a significantly faster rate than did girls' during elementary school, and by the end of formal schooling, boys' levels of school liking were significantly lower than girls'.

School liking was also linked with specific victimization trajectories. At the start of school, the children who liked school least were those who reported the highest levels of victimization. Early victims had the lowest levels of school liking in kindergarten and, as a group, differed significantly from their counterparts in the low group. For children in the high-chronic group, school liking began low (although not significantly so, relative to the low group) and, unlike children in the early group, remained low throughout their school careers. By Grade 12, the high-chronic group's school liking scores were significantly lower than the low group. These findings are consistent with the view that children's dislike of school is partially rooted in painful peer experiences in that context, and that these experiences can take a lasting toll on children's emotional engagement. Support was also found, however, for the recovery hypothesis in that, by Grade 12, school liking did not differ significantly for early versus low victims.

School avoidance was conceptualized as an indicator of children's motivation to evade the school context. Low to modest levels of school avoidance were exhibited by most students across time with the exceptions of chronic and early victims. Moreover, support was found for both the chronic stress hypothesis and recovery hypothesis. Specifically, consistent with a chronic stress hypothesis, school avoidance increased for high-chronic victims to the point that, by Grade 12, it was significantly higher than levels exhibited by low victims. Consistent with the recovery hypothesis, for early victims, school avoidance tendencies decreased to levels similar to low victims, reflecting reduced risk for peer victimization.

In kindergarten, however, it was the early victim group and not the chronic victim group that manifested significantly higher levels of school avoidance relative to low victims. This finding was unexpected because, at this point in their schooling, children in both groups were reporting similar levels of victimization. This discrepancy might be attributable to factors that were not assessed in this study (e.g., between group differences in child temperament, behavior, family circumstances) and raises the possibility that children who are destined to escape victimization differ in important ways from those that are not. Data from this study suggests that early victims were academically more prepared and engaged in kindergarten than were chronic victims. Such children, because of their greater investment in school (i.e., higher engagement, achievement), may have had stronger initial reactions to victimization (i.e., higher avoidance responses), but greater resources for overcoming victimization and its effects in the long run.

Independent participation indexed children's propensity to take initiative toward classroom activities. Girls exhibited higher levels of this form of behavioral engagement relative to boys, and, normatively, modest albeit nonsignificant declines in independent participation occurred across the course of formal schooling. When examined by trajectory groups, this aspect of school engagement was significantly lower for the high-chronic and moderate emerging subtypes. Specifically, compared to low victims, children in the high-chronic and moderate-emerging victimization groups not only had significantly lower levels of independent participation in kindergarten, but also retained these positions throughout their school careers.

Although lower independent participation in kindergarten was hypothesized for the high-chronic group, this difference was not expected for the moderate-emerging subtype until later grades (i.e., paralleling the pattern of increasing victimization). Thus, it was surprising that this group's participation trajectory during the early grades—a period during which its members reported only moderate victimization—resembled that of the chronic group. This finding raises the possibility that low classroom participation may be not only a consequence of victimization, but also an initial risk factor. That is, children in the moderate emerging group may have differed from their chronic counterparts in ways that made them not only less engaged in school but also more vulnerable to victimization as they matured. Such a profile, for example, might be manifested by passive or withdrawn children. Children with these propensities are likely to have persistent difficulties with classroom participation, and evidence suggests that their risk for peer victimization increases as they approach preadolescence (Younger, Gentile, & Burgess, 1993). Of course, these interpretations are speculative, and further research is needed to understand why some children's risk for victimization increases as they progress through school.

Academic self-perceptions. Support was found for the hypothesis that victimized children tend to have lower estimations of their academic competence. High-chronic victims' estimates of their academic competence were significantly lower than those of low victims beginning in kindergarten and remained this way across their entire school careers. Children who became more victimized over time (i.e., moderate-emerging victims) were not inclined to see themselves as less academically competent in kindergarten, but did so by Grade 12, consistent with chronic stress perspectives. These results support the view that victimized children are less likely than their nonvictimized counterparts to receive support (e.g., peer affirmation) or have peer-mediated classroom experiences that contribute to their sense of academic competence.

Achievement. Findings were consistent with the conclusion that peer victimization is associated with lower achievement (Nakamoto & Schwartz, 2010). However, evidence from this investigation provides a more comprehensive analysis of the temporal patterning of this relation across the formal school years, and implies that the strength of this association varies not only with victimization trajectories, but also by type of achievement.

For math achievement, the norm was for standard scores to rise during the grade- and middle-school years and then decline during high school. However, the rates and levels of this facet of achievement varied significantly by trajectory subtypes. Two principal patterns of association merit consideration. First, the evidence lent support to the view that any form of peer victimization disrupts children's mathematics achievement, particularly during earlier or foundational years of schooling. For children in all three victimization subtypes (early, moderate-emerging, and high-chronic), growth in math achievement from Grades 2 through 6 was significantly slower than the rate

observed for low victims. Second, the lowest levels of mathematics achievement were linked with chronic and emergent (increasing) victimization patterns. Consistent with the chronic stress hypothesis, children in the high-chronic subtype had math achievement that was significantly lower than low victims, and this difference was apparent from the beginning to the end of formal schooling. Those whose victimization began at moderate levels but increased over time had significantly lower math achievement by Grade 12.

Standard scores for reading exhibited, on the norm, more continuity across grades than did math scores and, when contrasted for victimization subtypes, stronger differences at earlier rather than later grades. As with math achievement, membership in the chronic and emergent (increasing) victimization subtypes was associated with significantly lower reading achievement in Grade 2 (compared to low victims). However, in contrast to math achievement, the children in these groups no longer differed from their low-victim counterparts at Grade 12. This was the only instance in which chronic victimization was not associated with lasting academic difficulties.

Limitations and Future Directions

Several limitations of past research were addressed in this study by mapping peer victimization's prevalence and normative (mean) trends, trajectory subtypes (classes), and academic linkages across the entire period of formal schooling. These strengths, however, were accompanied by certain limitations that should be considered in the context of the study's findings. First, the sampled school environments, particularly those incorporating transitions from self-contained classrooms (primary schools) to school environments where students attend classes with different teachers and students (middle and secondary schools), could have influenced the identified victimization trajectory subtypes and their academic associations. Although such influences were not in evidence (e.g., accounting for transition timing in analyses failed to support this premise), it remains important to ascertain whether the reported findings generalize to other types of school or cultural contexts. Second, although findings were consistent with the premise that peer victimization drives academic maladjustment, other hypotheses (e.g., academic difficulties engender victimization) merit consideration and should be evaluated in future studies. Third, although the cited advantages of self-report methods led us to utilize this strategy for select focal constructs, the use of additional methods, including multimethod assessment strategies, would strengthen future work. Fourth, it is possible that attrition during the later years of this study (i.e., high school) might have made some findings more representative or reliable for girls than for boys. However, by accounting for gender differences in each model, this was unlikely.

Summary and Conclusions

At the broadest level, the results imply that not only is peer victimization negatively associated with achievement, but it is also inversely related to several forms of academic engagement and self-perceptions, all of which have been established as achievement precursors. As a potential determinant, it is conceivable that peer victimization's toll on children's achievement stems from its capacity to undermine children's school engagement. These interconnections, and potential mediated relations, warrant further investigative attention.

Support for the hypothesis that victimization is especially disruptive during the foundational period (K–5) was found for mathematics but not reading achievement. During Grades 2 through 6, membership in any of the victimization subtypes (early, moderate-emerging, and high-chronic) was associated with lesser growth in mathematics, relative to low victims. These findings raise the possibility that early mathematics learning, more so than reading, is hindered by peer victimization. If this supposition receives additional support (e.g., replication), then the processes that underlie such a relation merit explication. Victimization's role in disrupting early math achievement might, for example, be mediated through its effects on children's emotions (e.g., dysregulation due to anger, anxiety), mental states (e.g., poor concentration, reduced ability to perform mental manipulations), or school engagement (e.g., excessive absences).

The findings also help to clarify how specific, long-term victimization patterns are associated with children's school engagement and achievement. Consistent with a chronic stress hypothesis, severe and enduring victimization, best exemplified by the high-chronic trajectory subtype, was often related to lower—and typically prolonged—disparities in school engagement, academic self-perceptions, and achievement. In large part, results for the other victimization subtypes showed that, when children became more victimized over time, they tended to exhibit lower or declining scores on these same academic indicators, and when they became less victimized over time (i.e., early victims), they exhibited higher or increasing scores on these indicators (i.e., data suggestive of "recovery"). Overall, these findings support—but do not confirm—a victimization-as-cause perspective (Nakamoto & Schwartz, 2010; Schwartz et al., 2005) in which it is argued that academic debilities are partially rooted in painful experiences that schoolmates perpetrate on children.

References

Biggs, B. K., Vernberg, E., Little, T., Dill, E. J., Fonagy, P., & Twemlow, S. W. (2010). Peer victimization trajectories and their association with children's affect in late elementary school. *International Journal of Behavioral Development, 34,* 136–146. http://dx.doi.org/10.1177/0165025409348560

Birch, S. H., & Ladd, G. W. (1997). The teacher–child relationship and children's early school adjustment. *Journal of School Psychology, 35,* 61–79. http://dx.doi.org/10.1016/S0022-4405(96)00029-5

Boivin, M., Petitclerc, A., Feng, B., & Barker, E. D. (2010). The developmental trajectories of peer victimization in middle to late childhood and the changing nature of their behavioral correlates. *Merrill-Palmer Quarterly, 56,* 231–260. http://dx.doi.org/10.1353/mpq.0.0050

Buhs, E., & Ladd, G. W. (2001). Peer rejection as antecedent of young children's school adjustment: An examination of mediating processes. *Developmental Psychology, 37,* 550–560. http://dx.doi.org/10.1037/0012-1649.37.4.550

Buhs, E. S., Ladd, G. W., & Herald, S. (2006). Peer exclusion and victimization: Processes that mediate the relation between peer group rejection and children's classroom engagement and achievement? *Journal of Educational Psychology, 98,* 1–13. http://dx.doi.org/10.1037/0022-0663.98.1.1

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255. http://dx.doi.org/10.1207/S15328007SEM0902

Delfabbro, P., Winefield, P., Trainor, S., Dollard, M., Anderson, S., Metzer, J., & Hammarstrom, A. (2006). Peer and teacher bullying/victimization of South Australian secondary school students: Prevalence and psychosocial profiles. *British Journal of Educational Psychology, 76,* 71–90.

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: Guilford Press.

Ettekal, I., Kochenderfer-Ladd, B., & Ladd, G. W. (2015). A synthesis of person- and relational-level factors that influence bullying and bystanding behaviors: Toward an integrative framework. *Aggression and Violent Behavior, 23,* 75–86. http://dx.doi.org/10.1016/j.avb.2015.05.011

Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research, 59,* 117–142. http://dx.doi.org/10.3102/00346543059002117

Finn, J. D. (1993). *School engagement and students at risk.* Washington DC: Department of Education, National Center for Educational Statistics (ERIC Document Reproduction Service No. ED 362 322).

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74,* 59–109. http://dx.doi.org/10.3102/00346543074001059

Furlong, M. J., Sharkey, J. D., Felix, E. D., Tanigawa, D., & Green, J. G. (2010). Bullying assessment: A call for increased precision of self-reporting procedures. In S. R. Jimerson, S. M. Swearer, & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 329–345). New York, NY: Routledge.

Goldbaum, S., Craig, W. M., Pepler, D., & Connolly, J. (2003). Developmental trajectories of victimization: Identifying risk and protective factors. *Journal of Applied School Psychology, 19,* 139–156. http://dx.doi.org/10.1300/J008v19n02_09

Harter, S. (1982). The perceived competence scale for children. *Child Development, 53,* 89–97. http://dx.doi.org/10.2307/1129640

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Hughes, S. (1987). Metropolitan Readiness Tests: 1986 Edition. In D. H. Keyser & R. C. Sweetland (Eds.), *Test critiques* (Vol. 6, pp. 341–349). Kansas City, MO: Test Corporation of America.

Juvonen, J., & Graham, S. (2014). Bullying in schools: The power of bullies and the plight of victims. *Annual Review of Psychology, 65,* 159–185. http://dx.doi.org/10.1146/annurev-psych-010213-115030

Kochel, K. P., Ladd, G. W., & Rudolph, K. D. (2012). Longitudinal associations among youth depressive symptoms, peer victimization, and low peer acceptance: An interpersonal process perspective. *Child Development, 83,* 637–650.

Kochenderfer, B. J., & Ladd, G. W. (1996). Peer victimization: Cause or consequence of school maladjustment? *Child Development, 67,* 1305–1317. http://dx.doi.org/10.2307/1131701

Kochenderfer-Ladd, B., & Wardrop, J. L. (2001). Chronicity and instability of children's peer victimization experiences as predictors of loneliness and social satisfaction trajectories. *Child Development, 72,* 134–151. http://dx.doi.org/10.1111/1467-8624.00270

Ladd, G. W. (1990). Having friends, keeping friends, making friends, and being liked by peers in the classroom: Predictors of children's early school adjustment? *Child Development, 61,* 1081–1100. http://dx.doi.org/10.2307/1130877

Ladd, G. W., Birch, S. H., & Buhs, E. S. (1999). Children's social and scholastic lives in kindergarten: Related spheres of influence? *Child Development, 70,* 1373–1400. http://dx.doi.org/10.1111/1467-8624.00101

Ladd, G. W., Buhs, E., & Seid, M. (2000). Children's initial sentiments about kindergarten: Is school liking an antecedent of early classroom participation and achievement? *Merrill-Palmer Quarterly, 46,* 255–279.

Ladd, G. W., & Dinella, L. M. (2009). Continuity and change in early school engagement: Predictive of children's achievement trajectories from first to eighth grade? *Journal of Educational Psychology, 101,* 190–206. http://dx.doi.org/10.1037/a0013153

Ladd, G. W., Herald-Brown, S. L., & Kochel, K. P. (2009). Peers and motivation. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 323–348). London, England: Routledge.

Ladd, G. W., Herald-Brown, S. L., & Reiser, M. (2008). Does chronic classroom peer rejection predict the development of children's classroom participation during the grade school years? *Child Development, 79,* 1001–1015. http://dx.doi.org/10.1111/j.1467-8624.2008.01172.x

Ladd, G. W., Kochenderfer, B. J., & Coleman, C. C. (1996). Friendship quality as a predictor of young children's early school adjustment. *Child Development, 67,* 1103–1118. http://dx.doi.org/10.2307/1131882

Ladd, G. W., & Kochenderfer-Ladd, B. (2002). Identifying victims of peer aggression from early to middle childhood: Analysis of cross-informant data for concordance, estimation of relational adjustment, prevalence of victimization, and characteristics of identified victims. *Psychological Assessment, 14,* 74–96. http://dx.doi.org/10.1037/1040-3590.14.1.74

Ladd, G. W., & Price, J. M. (1987). Predicting children's social and school adjustment following the transition from preschool to kindergarten. *Child Development, 58,* 1168–1189. http://dx.doi.org/10.2307/1130613

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 346–368). Newbury Park, CA: Sage. http://dx.doi.org/10.4135/9781412986311.n19

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.

Nakamoto, J., & Schwartz, D. (2010). Is peer victimization associated with academic achievement? A meta-analytic review. *Social Development, 19,* 221–242. http://dx.doi.org/10.1111/j.1467-9507.2009.00539.x

Nylund, K., Bellmore, A., Nishina, A., & Graham, S. (2007). Subtypes, severity, and structural stability of peer victimization: What does latent class analysis say? *Child Development, 78,* 1706–1722. http://dx.doi.org/10.1111/j.1467-8624.2007.01097.x

Olweus, D. (1999). Sweden. In P. K. Smith, Y. Morita, J. Junger-Tas, D. Olweus, R. Catalano, & P. Slee (Eds.), *The nature of school bullying: A cross-national perspective* (pp. 28–48). New York, NY: Routledge.

Olweus, D. (2010). Understanding and researching bullying: Some critical issues. In S. R. Jimerson, S. M. Swearer, & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 9–33). New York, NY: Routledge.

Paquette, J. A., & Underwood, M. K. (1999). Gender differences in young adolescent's experiences of peer victimization: Social and physical aggression. *Merrill-Palmer Quarterly, 45,* 242–266.

Ram, N., & Grimm, K. J. (2009). Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development, 33,* 565–576. http://dx.doi.org/10.1177/0165025409343765

Reavis, R. D., Keane, S. P., & Calkins, S. D. (2010). Trajectories of peer victimization: The role of multiple relationships. *Merrill-Palmer Quarterly, 56,* 303–332. http://dx.doi.org/10.1353/mpq.0.0055

Rohde, P., Lewinsohn, P. M., & Seeley, J. R. (1990). Are people changed by the experience of having an episode of depression? A further test of the scar hypothesis. *Journal of Abnormal Psychology, 99,* 264–271. http://dx.doi.org/10.1037/0021-843X.99.3.264

Rudolph, K. D., Troop-Gordon, W., Hessel, E. T., & Schmidt, J. D. (2011). A latent growth curve analysis of early and increasing peer victimization as predictors of mental health across elementary school. *Journal of*

Clinical Child and Adolescent Psychology, 40, 111–122. http://dx.doi.org/10.1080/15374416.2011.533413

Scholte, R. H. J., Engels, R. C., Overbeek, G., de Kemp, R. A. T., & Haselager, G. J. T. (2007). Stability in bullying and victimization and its association with social adjustment in childhood and adolescence. *Journal of Abnormal Child Psychology, 35,* 217–228. http://dx.doi.org/10.1007/s10802-006-9074-3

Schwartz, D., Gorman, A. H., Nakamoto, J., & Toblin, R. L. (2005). Victimization in the peer group and children's academic functioning. *Journal of Educational Psychology, 97,* 425–435. http://dx.doi.org/10.1037/0022-0663.97.3.425

Seals, D., & Young, J. (2003). Bullying and peer victimization: Prevalence and relationship to gender, grade level, ethnicity, self-esteem and depression. *Adolescence, 38,* 735–738.

Skinner, E. A., Wellborn, J. G., & Connell, J. P. (1990). What it takes to do well in school and whether I've got it: A process model of perceived control and children's engagement and achievement in school. *Journal of Educational Psychology, 82,* 22–32. http://dx.doi.org/10.1037/0022-0663.82.1.22

Snyder, J., Brooker, M., Patrick, M. R., Snyder, A., Schrepferman, L., & Stoolmiller, M. (2003). Observed peer victimization during early elementary school: Continuity, growth, and relation to risk for child antisocial and depressive behavior. *Child Development, 74,* 1881–1898.

Storch, E. A., Brassard, M. R., & Masia-Warner, C. L. (2003). The relationship of peer victimization to social anxiety and loneliness in adolescence. *Child Study Journal, 33,* 1–18.

Sullivan, T. N., Farrell, A. D., & Kliewer, W. (2006). Peer victimization in early adolescence: Association between physical and relational victimization and drug use, aggression, and delinquent behaviors among urban middle school students. *Development and Psychopathology, 18,* 119–137.

Sumter, S. R., Baumgartner, S. E., Valkenburg, P. M., & Peter, J. (2012). Developmental trajectories of peer victimization: Off-line and online experiences during adolescence. *Journal of Adolescent Health, 50,* 607–613. http://dx.doi.org/10.1016/j.jadohealth.2011.10.251

Thijs, J., & Verkuyten, M. (2008). Peer victimization and academic achievement in a multiethnic sample: The role of perceived academic self-efficacy. *Journal of Educational Psychology, 100,* 754–764. http://dx.doi.org/10.1037/a0013155

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in a growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Greenwich, CT: Information Age.

Troop-Gordon, W., & Ladd, G. W. (2005). Trajectories of peer victimization and perceptions of the self and schoolmates: Precursors to internalizing and externalizing problems. *Child Development, 76,* 1072–1091.

Wilkinson, G. S. (1993). *The wide range achievement test.* Wilmington, DE: Jastak Associates.

Wolke, D., Woods, S., Stanford, K., & Schulz, H. (2001). Bullying and victimization of primary school children in England and Germany: Prevalence and school factors. *British Journal of Psychology, 92,* 673–696.

Wu, W., West, S. G., & Taylor, A. B. (2009). Evaluating model fit for growth curve models: Integration of fit indices from SEM and MLM frameworks. *Psychological Methods, 14,* 183–201. http://dx.doi.org/10.1037/a0015858

Younger, A. J., Gentile, C., & Burgess, K. (1993). Children's perceptions of social withdrawal: Changes across age. In K. H. Rubin & J. B. Asendorpf (Eds.), *Social withdrawal, inhibition and shyness in childhood* (pp. 215–235). Hillsdale, NJ: Erlbaum.

Received December 2, 2015
Revision received November 16, 2016
Accepted November 18, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 6, 842–854

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000174

Short- and Long-Term Effects of Over-Reporting of Grades on Academic Self-Concept and Achievement

Fabio Sticca and Thomas Goetz
University of Konstanz and Thurgau University of Teacher Education

Ulrike E. Nett
University of Ulm

Kyle Hubbard
McGill University

Ludwig Haag
University of Bayreuth

This study examined the short- and long-term effects of self-enhancement (i.e., overreporting of academic grades) on academic self-concept and academic achievement. A total of 916, 719, and 647 students participated in the first, second, and third waves of assessment, respectively (mean age at T1 = 15.6 years). At each assessment, students reported their last midterm grades and their self-concepts in mathematics, German, English, and French. Actual midterm grades were obtained from the school administrations. Results showed that self-enhancement was positively associated with self-concept in the short term. However, in the long term, self-enhancement was directly associated with stronger decreases in self-concept and indirectly with stronger decreases in achievement that were linked to inflated self-concepts. Implications for research and educational practice are discussed.

Keywords: self-enhancement, academic self-concept, academic achievement, longitudinal data

A number of studies from different research fields have shown that there is a tendency to portray oneself as above average with respect to many individual characteristics such as driving skills (Svenson, 1981), relationship quality (Rusbult, Van Lange, Wildschut, Yovetich, & Verette, 2000), well-being (Goetz, Ehret, Jullien, & Hall, 2006; Wojcik & Ditto, 2014), and intelligence (Brown, 2012). This tendency is generally known as the *better-than-average effect* and is motivated by self-enhancement mechanisms (Alicke, 1985; Brown, 1986). Sedikides and Strube (1997) defined self-enhancement as "both the attempts to increase the positivity of one's self-concept (i.e., self-advancing) and attempts to diminish the negativity of one's self-concept" (i.e., self-protecting; p. 147). Self-enhancement takes on very different forms that have been described along several bipolar dimensions (Sedikides & Gregg, 2008) such as *public versus private* (i.e., toward others vs. oneself), *candid versus tactical* (i.e., based on opportunity or planned), and *relevant versus irrelevant* (i.e., in domains that are relevant or irrelevant for one's self-evaluation).

The conceptual opposite of self-enhancement is self-handicapping. Self-handicapping involves erecting performance-inhibiting barriers to either protect self-image following failure (i.e., discounting) or enhance one's self-image following success in very challenging situations (i.e., augmenting; Rhodewalt, Morf, Hazlett, & Fairfield, 1991). While self-enhancement is expected to increase long-term engagement and persistence in an activity (Taylor & Brown, 1994), self-handicapping might put students at risk of decreased performance in the long run (Schwinger, Wirthwein, Lemmer, & Steinmayr, 2014) even though they might be able to maintain a stable self-view in the short term (McCrea & Hirt, 2001), which shows that a short-term effect of self-handicapping might actually be self-enhancement (in one of its forms; i.e., maintaining a stable self-concept). Thus, both self-enhancement and self-handicapping are aimed at optimizing/maintaining one's self-view, but their long-term effects on achievement might be different. The present study examined the short- and long-term effects of self-enhancement on academic self-concept and achievement.

The *relevant versus irrelevant* dichotomy of self-enhancement highlights how self-enhancement is highest with respect to personal characteristics that individuals consider to be important (Brown, 2012). In the academic context, one important personal characteristic is academic achievement,[1] which we consider a latent construct typically manifested in numerical and letter grades. Achievement is often measured using exams, presentations, and other forms of academic assessment at various points through the semester or school year. Grades are a very salient and institutionalized form of feedback; thus, from the students' perspective,

This article was published Online First March 16, 2017.

Fabio Sticca and Thomas Goetz, Department of Empirical Educational Research, University of Konstanz and Department of Empirical Educational Research, Thurgau University of Teacher Education; Ulrike E. Nett, Institute of Psychology and Education, University of Ulm; Kyle Hubbard, Department of Educational and Counselling Psychology, McGill University; Ludwig Haag, Department of Education, University of Bayreuth.

Ulrike E. Nett is now at the Department of Empirical Educational Research, University of Augsburg.

Correspondence concerning this article should be addressed to Fabio Sticca, who is now at Institute of Education, University of Zurich, Kantonsschulstrasse 3, 8001 Zurich, Switzerland. E-mail: fabio.sticca@uzh.ch

[1] Unless otherwise indicated, the term "achievement" will be used to indicate academic achievement for the remainder of the article.

grades can be a meaningful source of information from which their achievement can be inferred (Pekrun, Hall, Goetz, & Perry, 2014). As such, grades can exert a direct influence on students' academic self-concept[2] (Niepel, Brunner, & Preckel, 2014; Preckel, Niepel, Schneider, & Brunner, 2013) and self-evaluation (Crocker, Karpinski, Quinn, & Chase, 2003).

Given the subjective relevance of academic grades and the tendency to self-enhance when reporting on important personal characteristics (Brown, 2012), it is not surprising that a general propensity to overreport past grades has been observed in many studies. Kuncel, Credé, and Thomas (2005) reviewed the existing research on the inaccuracy of self-reported grade point averages (GPAs) and carried out a meta-analysis of 37 independent samples encompassing a total of 60,926 individuals. The authors found that students tend to overreport their GPA and that the percentage of students overreporting their GPA is three to four times higher than the percentage who underreport. Studies on the validity of self-reported grades that were carried out after the meta-analysis by Kuncel et al. largely confirmed these findings (Dickhäuser & Plenter, 2005; Gramzow & Willard, 2006; Möller, Streblow, Pohlmann, & Köller, 2006; Schneider & Sparfeldt, 2016; Schwartz & Beaver, 2015; Shaw & Mattern, 2009; Sparfeldt, Buch, Rost, & Lehmann, 2008; Talento-Miller & Peyton, 2006). In particular, all studies but one (Shaw & Mattern, 2009) replicated the significant tendency to overreport grades.

One central particularity about the tendency to overreport grades is that it refers to *past* achievement rather than *future* achievement. Past achievement is usually known and can therefore be over-*reported*, while future achievement is unknown and can therefore only be over-*estimated*. Thus, when it come to evaluating one's performance, the temporal perspective is central. Willard and Gramzow (2008) found that the tendency for students to retrospectively report test scores that were higher than what they actually achieved becomes more marked with increasing temporal distance. On the other hand, overestimating *future* grades has also been conceptualized as a form of self-enhancement and was found to be negatively associated with academic achievement (Buckelew, Byrd, Key, Thornton, & Merwin, 2013). Self-enhancement that refers to future events has also been studied in research on *calibration of self-concept* (Alexander, 2013). Calibration of self-concept is therefore a specific form of self-enhancement that describes the match between one's perception of ability (i.e., self-concept) and one's actual ability (e.g., task performance; Bol, Hacker, O'Shea, & Allen, 2005): Students can either accurately (i.e., high calibration) or inaccurately (i.e., low calibration) judge their ability. Those students who inaccurately judge their ability can be further divided into under- and overconfident students (as opposed to under- and overreporting of past grades). In sum, overreporting past grades and overestimating future grades (i.e., overconfident calibration) differ in that past grades simply need to be remembered and reported, while a precise calibration of one's self concept is needed to be able to adequately predict one's future performance (Chiu & Klassen, 2010). In the present study, we focused on overreporting of past grades.

Although there is a paucity of research that has examined the psychological mechanisms that underlie the tendency to overreport academic grades, there is evidence suggesting that it represents a self-enhancement mechanism. Dickhäuser and Plenter (2005) found that overreporting of grades was positively correlated with academic self-concept in the domain of mathematics. The authors suggested that this might be indicative of a self-enhancement mechanism but did not expand on this suggestion. Similarly, Gramzow and Willard (2006) showed that GPA exaggeration is associated with self-enhancement, and Willard and Gramzow (2008) considered the tendency to report test scores as higher than they were to be a form of self-enhancement. Furthermore, it was proposed that such enhanced self-reports might be explained by the need to perceive the self as constantly improving (Ross & Wilson, 2003). Schwartz and Beaver (2015) argued that this need might be motivated by the pressure to obtain good grades that is experienced in school, family, and peer contexts. Considering that the tendency to overreport grades has been observed in many studies where participants knew that no link between their responses and their identity was possible (i.e., when completing an anonymous self-report questionnaire), it can be assumed that this specific form of self-enhancement might be private rather than public. Furthermore, Sedikides and Gregg (2008) discussed that self-enhancement can also manifest itself in less obvious ways than individuals explicitly stating that they believe to be above average with respect to some task. Reporting about *past* grades might be one of these less obvious manifestations of self-enhancement, as students may not be explicitly aware that their report will be used to determine if they self-enhanced (i.e., implicit assessment). Thus, it might be assumed that overreporting past grades might be candid rather than tactical. While overreporting grades has been interpreted as an indicator of self-enhancement, it is unclear if underreporting grades might be interpreted as an indicator of self-handicapping. Thus, we cannot safely assume that overreporting and underreporting are two ends of a continuum in terms of effects on self-concept and achievement. Accordingly, we will focus solely on overreporting as an indicator of self-enhancement. In light of these studies showing that reporting past grades and estimating future grades involves self-enhancement mechanisms, questions arise concerning whether this specific form of self-enhancement is adaptive or maladaptive in terms of both self-concept and achievement and whether its effects differ in the short and long term. The present study addresses these questions with a focus on the short- and long-term effects of overreporting *past* grades (i.e., self-enhancement) on self-concept and achievement.

Self-enhancement has been shown to be adaptive in the short term because it is positively associated with self-concept and achievement (Dickhäuser & Plenter, 2005; Kuncel et al., 2005), positively associated with self-esteem (i.e., the global component of self-concept) and well-being (Robins & Beer, 2001), and negatively associated with depressive symptoms (Noble, Heath, & Toste, 2011). Regarding calibration, Chiu and Klassen (2010) found that better calibration of mathematics self-concept was associated with higher mathematics self-concept and achievement. Moreover, the authors found that students who overestimated their mathematics self-concept had lower mathematics achievement. In sum, the short-term effects of self-enhancement seem to be adaptive in terms of academic self-concept and achievement, while those of overconfident calibration seem to be maladaptive.

Results on the long-term effects of self-enhancement obtained thus far are controversial. From a theoretical point of view, self-enhancing perceptions are assumed to increase motivation, persistence, and

[2] Similarly, the term "self-concept" will be used to indicate academic self-concept for the remainder of the article.

performance (Taylor & Brown, 1988, 1994). However, in a longitudinal study on the short- and long-term effects of positive illusions (i.e., a form of self-enhancement), Robins and Beer (2001) found that self-enhancement led to a number of maladaptive developments in the long term such as decreases in self-esteem and well-being, as well as increased disengagement from the school context across several years. However, Robins and Beer found no direct cross-sectional or longitudinal associations between self-enhancement and academic performance or graduation rates. Vancouver and Kendall (2006) found that overestimation of one's ability might negatively affect preparation and lead to lower performance. Furthermore, Ackerman and Wolman (2007) discussed that people who believe they can outperform their peers might exhibit inflated self-concepts, which might in turn lead to less preparation and help-seeking, and poor performance (Stone & May, 2002).

Taken together, these findings suggest that the short-term effects of self-enhancement on self-concept and achievement are predominantly positive, which conforms to the definition of self-enhancement and its psychological mechanisms. As an example, students that self-enhance, be it consciously or unconsciously, publicly or privately, might feel better in that very moment (i.e., better self-concept). However, the long-term effects of self-enhancement were found to be predominantly negative, which might be due to inflated self-concepts (i.e., the short-term benefit of self-enhancement) that might put students at risk of less learning effort and, consequently, lower achievement. For instance, students that tend to self-enhance might feel more competent than they actually are (i.e., overestimated self-concept) with respect to a given task (e.g., an exam) and might therefore be less prone to show behaviors that are necessary to be able to perform well. Thus, what appears to be adaptive in the short term (i.e., better self-concept) might turn out to be maladaptive in the long term (i.e., lower achievement).

To date, research on the longitudinal interplay between self-enhancement, self-concept, and achievement is scarce. Previous research suggests that self-enhancement increases one's self-concept in the short term (Dickhäuser & Plenter, 2005; Sedikides & Strube, 1997). Furthermore, existing evidence indicates that there is a positive and reciprocal longitudinal relation between self-concept and achievement within the same academic subject (Marsh, 1986; Marsh & Craven, 2006; Möller, Retelsdorf, Köller, & Marsh, 2011; Niepel et al., 2014). While the *bivariate* associations between self-enhancement and self-concept, and self-enhancement and achievement, were examined in a number of studies, there has yet to be a study that simultaneously examines the effects of self-enhancement on self-concept and achievement across high school using a longitudinal and *trivariate* approach. Since self-enhancement was found to be associated with self-concept, and self-concept was found to be associated with achievement, a trivariate approach is needed to explore how these three constructs work in concert. In this regard, it might be assumed that self-enhancement leads to a higher self-concept (Dickhäuser & Plenter, 2005) or self-esteem (Robins & Beer, 2001) in the short term. In turn, self-concept or self-esteem may lead to higher achievement as these constructs are positively and reciprocally associated with each other both cross-sectionally and longitudinally within the same academic subject (Marsh, 1986; Marsh & Craven, 2006; Möller et al., 2011; Niepel et al., 2014). Alternatively, an inflated self-concept or self-esteem may lead to a decrease in achievement, possibly resulting from less effort invested in learning and achievement-striving (Robins & Beer, 2001; Stone & May, 2002; Svanum & Bigatti, 2006).

Thus far, only Robins and Beer (2001) have examined all of these constructs simultaneously within a longitudinal framework (i.e., self-enhancement, self-concept, and achievement), although they focused on self-esteem instead of self-concept. In addition to utilizing more sophisticated statistical methods, namely latent growth modeling (LGM), their study was the first to adopt a longitudinal approach to examine long-term effects of self-enhancement while at the same time using an external criterion (i.e., ability measured by SAT scores) to operationalize self-enhancement (i.e., difference between self-evaluated and actual ability). However, Robins and Beer examined *bivariate* longitudinal associations only. Therefore, the role of self-concept in the longitudinal association between self-enhancement and achievement has yet to be explored. As outlined above, self-enhancement, self-concept, and achievement are associated with each other. Thus, it is important to examine the longitudinal development of these three constructs in concert. If the associations between constructs are only studied in a bivariate framework, more complex trivariate associations (e.g., indirect effects) might remain undetected, thereby leading to incomplete conclusions. By taking a trivariate approach, we aim at overcoming this methodological limitation and expanding our knowledge on the longitudinal interplay between self-enhancement, self-concept, and achievement.

The purpose of the present study was to examine the longitudinal interplay between self-enhancement, self-concept, and achievement using a trivariate approach. Our first aim was to replicate cross-sectional results pertaining to the association between self-enhancement, self-concept, and achievement. In this regard, we hypothesized (a) that higher levels of self-enhancement would be cross-sectionally associated with higher scores of self-concept (Dickhäuser & Plenter, 2005; Robins & Beer, 2001), (b) that higher levels of self-enhancement would be associated with higher achievement (Kuncel et al., 2005), and (c) that higher levels of self-concept would be associated with higher levels of achievement within the same academic subject (Marsh, 1986; Marsh & Craven, 2006; Niepel et al., 2014).

Our second aim was to explore the long-term effects of self-enhancement on self-concept and achievement. In line with results obtained by Robins and Beer (2001), we hypothesized that self-enhancement would be negatively associated with the development of self-concept. Based on previous finding also from Robins and Beer (2001), we did not expected to find a significant direct association between the initial level of self-enhancement and the development of achievement; thus, no specific hypothesis was constructed in this regard. Additionally, we hypothesized that there would be a reciprocal and positive longitudinal association between self-concept and achievement (Marsh, 1986; Marsh & Craven, 2006; Möller et al., 2011; Niepel et al., 2014). Finally, we explored potential indirect long-term effects of self-enhancement on achievement that were mediated by self-concept.

Method

Sample and Procedure

The present study was conducted in the German-speaking part of Switzerland. A total of three assessments were carried out in the

spring of 2012 (T1), 2013 (T2), and 2014 (T3). The timing of the assessments was designed so that the entire period of upper-track school in Switzerland was covered (known as Gymnasium schools in the Swiss-based state school system). Since most students attend the same school for these three years, this design element also ensured that the academic context was stable over time. As most students move to vocational or tertiary education after Gymnasium, the Gymnasium years are a crucial period in the development of motivational constructs such as self-concept, which is highly relevant for the transition to higher education. Finally, 1-year intervals are typically chosen for the study of long-term developments (e.g., Robins & Beer, 2001).

From all German-speaking upper-track schools in Switzerland where the four academic subjects of mathematics, German, English, and French were taught in Grades 9 to 11, eight Gymnasiums were randomly selected for participation in the present study. All students in the 45 Grade 9 classrooms from these eight schools were eligible to participate. A total of 916 students participated in the first assessment (56.1% female; mean age 15.6 years, SD = .63), 719 participated in the second assessment (55.5% female; mean age 16.6 years, SD = .63), and 647 participated in the third assessment (55.3% female; mean age 17.7 years, SD = .75). Attrition was mainly due to one school dropping out of the study after the first assessment (n = 146), to students leaving the school they were initially assessed at, or to students being absent during data collection. To avoid a substantial drop in statistical power due to the reduction in sample size, 42 students were additionally recruited at T2 and 38 students were additionally recruited at T3. A subsample of 571 (57.4%) students participated in all three assessments, while 145 (14.6%) participated in two assessments, and 280 (28%) participated in only one assessment. In sum, a total of 996 students participated in at least one measurement occasion of the present study.

A total of 90.7% of the participants were born in Switzerland, while 6.2% were born in other European countries. Regarding the participants' parents' nationality, the respective percentages were 68.8% and 19.6% for participants' mothers and 71.0% and 19.5% for their fathers. A total of 87.1% of the students spoke German at home, while 1.0% spoke French and 0.8% spoke Italian. Among those participants not speaking a national language at home, the three most common languages were Albanian (1.4%), Tamil (1.1%), and Turkish (1.0%). Regarding parents' education, 31.6% of the participants' mothers and 46.2% of their fathers held a university or college degree. Of those parents without a university degree, 47.6% of mothers and 40.8% of fathers held a vocational college degree, and 12.6% of mothers and 11.6% of fathers had a high-school diploma. 0.5% of the participants' parents had not completed high school.

Assessments were carried out in the classrooms during a single, 45-min lesson using a paper and pencil questionnaire. Before the first assessment, participants were informed that participation in the study was voluntary and that they could discontinue their involvement at any time without any negative consequences. Furthermore, all parents or guardians were informed about the study, its aims, and its procedures. The heads of schools and the teachers who taught in the classes from which the participants were drawn approved the study protocol. Every participant was given a personal identification number and was asked to write it on their questionnaire before beginning. After the data were collected and entered, all identifiers linking participants to their data were deleted. Thus, analyses were conducted on depersonalized data. After each assessment, participants were compensated with a small gift, such as chocolate, and entry into a prize drawing to win an Apple iPod.

Regarding the sequence of assessment for the three constructs of interest, it must be noted that while actual grades were given to the students in December, students' self-reported grades, self-enhancement, and academic self-concept, which corresponded with their December grades, were assessed in the spring of the subsequent year.

Study Measures

Demographic variables. Participants' gender and age were obtained via self-report at each assessment.

Actual academic achievement. Each student's midyear grades (i.e., grades obtained in December of the previous year, roughly four months before the assessments at T1, T2, and T3) in mathematics, German, English, and French were provided by the respective school administrations at each assessment and were linked to the individual data using anonymous identification codes. In Switzerland, grades range from 1 (*insufficient*) to 6 (*excellent*), with 4 being the threshold for a sufficient grade. Half grades (e.g., 4.5) are also common in Switzerland. Grades are generally determined by the results that students obtain in their exams across a term. The exam formats vary as a function of the academic subject. For instance, mathematics exams usually consist of solving mathematical problems, while compositions, presentations, and vocabulary tests are common in linguistics courses. In the foreign languages (i.e., English and French in the present study), translations are also used as a form of exam. Table 1 shows the mean scores and standard deviations of actual grades at each assessment and for each academic subject.

Self-reported academic achievement. At each assessment, participants were asked to report their last midyear grades (i.e., grades obtained in December of the previous year, roughly four months before the assessments at T1, T2, and T3) in mathematics, German, English, and French classes.

Self-enhancement. In line with recommendations by Robins and Beer (2001), we adopted an external criterion to operationalize self-enhancement, which was defined as the amount that students overreported their academic grades. This operationalization comes with the advantage that the resulting measure is less biased than other measures for self-enhancement (e.g., estimating one's performance relative to the perceived performance of others). Furthermore, the resulting measure is continuous instead of categorical (i.e., better, equal, or worse than others' performance).

Initially, actual academic grades were subtracted from the self-reported academic grades. This operation resulted in students with negative scores (i.e., the self-reported grade was lower than the actual grade, which will be labeled *underreporting*), null scores (i.e., accurate reporting), and positive scores (i.e., the self-reported grade was higher than the actual grade, which will be labeled *overreporting*). The difference between the actual grade and the self-reported grade does *not* represent a clear operationalization of overreporting, as it also encompasses underreporting. As outlined above, we cannot safely assume that underreporting and overreporting are two ends of a continuum that can be labeled self-enhancement. If we did, then

Table 1
Descriptive Statistics for Self-Reported Academic Achievement, Actual Academic Achievement, Self-Enhancement, and Academic Self-Concept in Mathematics, German, English, and French

Variable	T1 (2012) *M*	T1 (2012) *SD*	T2 (2013) *M*	T2 (2013) *SD*	T3 (2014) *M*	T3 (2014) *SD*
Self-reported academic Achievement						
Mathematics	4.57	.74	4.51	.82	4.50	.80
German	4.73	.46	4.70	.57	4.75	.57
English	4.71	.61	4.71	.65	4.72	.62
French	4.59	.69	4.55	.73	4.48	.73
Actual academic Achievement						
Mathematics	4.52	.75	4.44	.78	4.42	.82
German	4.67	.46	4.66	.52	4.72	.54
English	4.65	.62	4.67	.62	4.65	.62
French	4.54	.68	4.51	.70	4.44	.72
Self-enhancement						
Mathematics	.08	.23	.12	.30	.13	.29
German	.09	.22	.09	.22	.11	.24
English	.08	.23	.09	.22	.12	.28
French	.09	.25	.10	.28	.10	.28
Academic self-concept						
Mathematics	3.14	1.10	3.02	1.11	3.01	1.13
German	3.30	.89	3.20	.92	3.26	.97
English	3.41	1.05	3.39	1.04	2.76	1.21
French	3.17	1.09	3.07	1.13	2.90	1.11

Note. *M* = Mean; *SD* = Standard deviation; T1 = Time 1 assessment (N = 916); T2 = Time 2 assessment (N = 719); T3 = Time 3 assessment (N = 647).

underreporting would represent self-handicapping, which is not the opposite of self-enhancement in terms of its effects on self-concept and achievement (Sedikides & Gregg, 2008). In order to actually operationalize self-enhancement in the context of this study, participants that underreported their grade were given a score of 0 on self-enhancement. This was done for each academic subject. Notably, this transformation led to a decrease in the variance of self-enhancement, which resulted in decreased correlations of self-enhancement with achievement and self-concept, and, therefore, to a more conservative analysis strategy. Table 1 displays the mean scores and standard deviations of self-enhancement at each assessment and for each academic subject.

Academic self-concept. The Self-Description Questionnaire (Marsh & O'Neill, 1984) was used to assess self-concept in mathematics, German, English, and French. The scale encompassed a total of three items: (1) I get good marks in [ACADEMIC SUBJECT]; (2) [ACADEMIC SUBJECT] is one of my best subjects; and (3) I have always done well in [ACADEMIC SUBJECT]. Response options consisted of a 5-point Likert scale (from 1 = *strongly disagree* to 5 = *strongly agree*). A mean score of the three items was computed for each academic subject separately and was used in the following analyses. The internal consistencies (Cronbach's alpha) of the mean scores across all subjects and within assessments were found to be between .84 and .91. Table 1 shows the mean scores and standard deviations referring to the summative scales divided by the number of scale items.

Data Analysis

The main aim of the present paper was to examine the longitudinal association between self-enhancement, self-concept, and achievement. Before addressing the main research question, unconditional multilevel models were used to assess the intraclass correlation (ICC) of self-enhancement across the four academic subjects. Individuals were modeled as level 1 units and classrooms were modeled as level 2 units. These analyses showed that the ICC of self-enhancement was .040, .024, and .039 at T1, T2, and T3, respectively. The respective ICCs for self-concept were .014, .013, and .037, while those for achievement were .042, .035, and .038. These results revealed that almost all of the variance in the three variables lay at the individual level, while hardly any variance lay at the class level (Heinrich & Lynn, 2001; Lee, 2000). Nevertheless, we did take the classroom level into account in order to address the dependence of observation within classrooms. This was achieved using the sandwich estimator.

The longitudinal interplay between self-enhancement, self-concept, and achievement was examined using a trivariate parallel process latent growth model (TPPLGM; King, Nguyen, Kosterman, Bailey, & Hawkins, 2012), which is an extension of the parallel process latent growth model (PPLGM; Chung, White, Hipwell, Stepp, & Loeber, 2010). This model allowed us to test whether latent growth parameters of one latent growth model (LGM) predicted those of another LGM.[3]

Our aim was to model a single TPPLGM in which it would be possible to examine the hypotheses while taking the four academic subjects into account as covariates. All variables were collected with respect to the four academic subjects. Therefore, the structure of the data was crossed, with student being nested in classes and academic subjects. Thus, it was necessary to restructure the dataset. More precisely, it was necessary to obtain only one variable for each one of the constructs of interest (e.g., self-concept) instead of four (e.g., mathematics self-concept, German self-concept, English self-concept, French self-concept). Accordingly, we decided to restructure the data so that every student would have four data rows, where the first row would contain the scores relative to mathematics, followed by a second, third, and fourth row containing the information relative to German, English, and French,

[3] If a linear development is assumed, each LGM will encompass an intercept and a slope, which describe the *intra*individual development across time. The intercept represents the initial score, while the slope describes how scores develop over time (i.e., increase vs. decrease). The intercept and the slope variances capture interindividual differences in intraindividual development, and therefore their correlation is highly informative. For instance, a positive correlation indicates that higher initial scores are associated with more positive slopes over time. It is important to note that the meaning of *more positive* depends on the mean slope. If the mean slope across all students is negative, a more *positive* slope indicates that the decrease is less pronounced and might even turn into an increase. If the mean slope is negative, a more *negative* slope indicates that the decrease becomes even more pronounced. The reverse rationale applies to positive mean slopes. Associations between intercepts and slopes can be examined within a construct as well as across multiple constructs. Moreover, one can also examine the correlations among multiple intercepts and among multiple slopes of different constructs. However, no causal interpretations are possible, as the intercept might be causally influenced by earlier events that were not included in the model. Accordingly, associations between latent growth parameters are usually modeled as correlations. For more information on LGMs, see Bollen and Curran (2005).

respectively. This operation multiplied the length of the dataset by four and reduced the number of variables to one for each construct (i.e., self-concept, actual grade, and self-enhancement). The effect of the academic subjects on the growth parameters was controlled for in the analyses. This strategy has been discussed and applied as a way to deal with crossed data structures (Goetz, Sticca, Pekrun, Murayama, & Elliot, 2016; Huang, 2016). With this data structure it was possible to model a single TPPLGM instead of four different TPPLGMs for the four academic subjects. Table 2 shows the zero-order correlations between all study variables. Correlations are reported separately for each academic subject.

Separate univariate LGMs were modeled for self-enhancement, self-concept, and achievement in order to assess their model fit. All LGMs were modeled as first-order LGMs using the three observed scores of the respective constructs to estimate a latent intercept and a latent slope (i.e., a linear development was assumed). Accordingly, the factor loadings from the latent intercept to the observed scores of absolute inaccuracy were all set to 1, while those of the latent slope were set to 0, 1, and 2 (Bollen & Curran, 2005). The three LGMs were then combined into a TPPLGM, and covariances between the latent growth parameters (i.e., intercepts and slopes) were modeled. The residual variances of the observed variables that were assessed at the same time point were also allowed to covary (e.g., self-enhancement at T1 and self-concept at T1). No further modifications were made to the TPPLGM. In the final step, we proceeded to extend the TPPLGM by including gender, age, and academic subject as time-invariant covariates. Analyses were performed using Mplus 7.11 (Muthén & Muthén, 2012). As the self-enhancement variables were non-normally distributed, robust maximum likelihood was used as an estimation algorithm.[4]

Participant attrition across the study was largely due to one school dropping out after the first assessment because of organizational issues that were unrelated to any of the variables under examination in the present study. Other less prominent causes of attrition were students leaving a school or being absent during data collection, which could be assumed to be unrelated to any of the variables under examination in the present study. Accordingly, it was assumed that data were missing at random, and the full information maximum likelihood (FIML) method was used to address missing data. To evaluate the extent to which the FIML procedure was appropriate for the longitudinal analyses at hand, we compared (a) the mean scores of self-concept, actual achievement, and self-reported achievement of students with complete data to (b) the same means scores that were obtained from the entire sample using the FIML method for the imputation of missing values. Results showed that means scores were almost identical for all pairs of means scores, suggesting that the FIML procedure was indeed well suited.

Results

Univariate Longitudinal Development of Self-Enhancement, Academic Self-Concept, and Academic Achievement

The univariate LGMs for self-enhancement, self-concept, and achievement were found to match the data well (see Table 3). Table 4 shows the mean scores and the standard deviations of the three LGMs. Note that these are not estimates gained from the univariate LGMs but from the TPPLGM without covariates, which were virtually equal to those obtained from the univariate LGMs. Regarding self-enhancement, we found that the initial level (i.e., intercept) was almost one tenth of a grade. This initial score was found to have significant variance. The change over time (i.e., the slope) of self-enhancement was found to be positive, but its variance was not significant. In other words, on average, students were found to self-enhance at the first assessment, and this tendency was found to increase over time. Students were found to differ in their initial level of self-enhancement, but the increase in self-enhancement from Grade 9 to Grade 11 was found to be the same for all students. As for self-concept, we found that the initial score was close to the middle of the scale and had a significant variance. Over time, self-concept was found to significantly decrease, and this decrease had a significant variance. Thus, students were found to differ regarding their initial level of self-concept and its development over time, with most students experiencing a decrease. Finally, the results for achievement showed that the initial score was found to be half a point above the threshold for a sufficient grade and to have a significant variance. The longitudinal trend in achievement was also found to be significantly negative and to have a significant variance. Accordingly, students were found to differ regarding their initial level of achievement and its development over time, with most students experiencing a decrease. In summary, from Grade 9 to Grade 11, self-enhancement was found to increase while self-concept and achievement were found to decrease (see Figure 1).

Cross-Sectional Relations Between Self-Enhancement, Academic Self-Concept, and Academic Achievement

The correlations between all latent growth parameters are reported in Table 4. Additionally, Figure 1 shows the standardized solution of the TPPLGM. Note that error covariances are not

[4] The distribution of self-enhancement was skewed. A potential modeling strategy that would address this complication is the two-part latent growth model (TP-LGM). Herein, data are split into a dichotomous part and a linear part. In the dichotomous part, scores are recoded into 0 and 1, where a score of 1 is given to those who self-enhanced, independently of the amount of self-enhancement, and a score of 0 is given to all others. In the continuous part, those who did not self-enhance are given a missing value, while all others retain their score. Thus, the dichotomous part describes the initial percentage of self-enhancers (i.e., intercept) and its change (i.e., slope). The continuous part describes the initial level of self-enhancement and its change. These two parts can then be joined into a parallel process model. The modeling of the dichotomous part requires one dimension of integration for each latent variable, resulting in two dimensions of integration in this specific case (i.e., intercept and slope). The model becomes computationally demanding when adding further processes to the model such as self-concept and achievement. In particular, the correlations of residuals from the same time point need to be modeled in order to not distort the correlations among the latent growth parameters. As correlations between linear and dichotomous indicators cannot be modeled, a latent variable has to be modeled. In the present analysis, this resulted in a total of five dimensions of integration. Even more computational demands arise when adding covariates to the model. Finally, indirect effects cannot be examined within this framework. In sum, the option to run a TP-LGM for self-enhancement and to integrate it into a TP-TPPLGM seemed to be impracticable in the present study. Therefore, a traditional approach with underlying assumptions of normality was taken.

Table 2
Zero-Order Correlations Between All Study Variables for Mathematics (Above the Diagonal in the Upper Half of the Table), German (Below the Diagonal in the Upper Half of the Table), English (Above the Diagonal in the Lower Half of the Table), and French (Below the Diagonal in the Lower Half of the Table)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Mathematics (above the diagonal) and German (below the diagonal)														
1. Sex (male)	1	−.03	.21**	.16**	.16**	.01	−.01	−.05	−.02	−.06	−.03	.03	.08*	−.01
2. Age (years)	−.03	1	.02	−.05	−.03	.04	−.02	−.01	.03	−.02	−.06	.07	.06	.10*
3. ACSC T1	−.19***	−.03	1	.73***	.69***	.74***	.55***	.51***	.71***	.50***	.47***	.04	.15***	.04
4. ACSC T2	−.22***	−.03	.64***	1	.75***	.59***	.71***	.56***	.55***	.69***	.55***	.09*	.12**	−.03
5. ACSC T3	−.25***	−.02	.61***	.74***	1	.56***	.61***	.75***	.57***	.59***	.71***	.01	.13***	.04
6. S-R Grade T1	−.17***	−.01	.62***	.42***	.44***	1	.60***	.58***	.93***	.59***	.56***	.13***	.06	.01
7. S-R Grade T2	−.19***	−.06	.32***	.53***	.51***	.44***	1	.65***	.57***	.86***	.67***	.09*	.35***	−.06
8. S-R Grade T3	−.23***	.04	.32***	.46***	.63***	.42***	.53***	1	.59***	.65***	.88***	.01	.08	.20***
9. Actual Grade T1	−.21***	.01	.55***	.39***	.42***	.85***	.43***	.42***	1	.58***	.58***	−.24***	.05	−.03
10. Actual Grade T2	−.23***	−.03	.36***	.54***	.47***	.44***	.76***	.50***	.47***	1	.66***	.06	−.18***	−.04
11 Actual Grade T3	−.24***	.02	.30***	.40***	.58***	.47***	.52***	.77***	.50***	.53***	1	−.02	.09*	−.29**
12. Self-Enhancement T1	.06	.01	.11***	.06	.04	.26***	.01	−.01	−.29***	−.06	−.04	1	.04	.07
13. Self-Enhancement T2	.02	−.06	.01	.06	.10*	.09*	.48***	.05	.04	−.21***	.02	.08*	1	−.05
14. Self-Enhancement T3	.04	.02	.06	.103*	.11**	−.04	.04	.39***	−.06	−.01	−.29***	.04	.08	1
English (above the diagonal) and French (below the diagonal)														
1. Male	1	−.03	−.07	−.07	.01	−.13***	−.16***	−.11**	−.16***	−.18***	−.13**	.08*	.01	.06
2. Age	−.03	1	−.02	−.07	.08	.01	−.05	−.08	−.03	−.03	−.12**	.06	.02	.08
3. ACSC T1	−.25***	−.08*	1	.79***	.06	.76***	.59***	.51***	.73***	.59***	.53***	.04	.05	−.02
4. ACSC T2	−.27***	−.09*	.77***	1	.08	.67***	.73***	.59***	.64***	.70***	.57***	.02	.12**	.01
5. ACSC T3	−.32***	−.10*	.72***	.82***	1	.04	.05	.02	.06	.03	−.03	−.05	.02	.07
6. S-R Grade T1	−.23***	−.05	.75***	.63***	.59***	1	.69***	.58***	.91***	.69***	.64***	.16***	.05	−.08
7. S-R Grade T2	−.26***	−.05	.60***	.75***	.71***	.64***	1	.64***	.69***	.86***	.65***	−.05	.34***	−.05
8. S-R Grade T3	−.36***	−.17***	.56***	.64***	.75***	.54***	.65***	1	.57***	.64***	.81***	.01	.03	.29***
9. Actual Grade T1	−.24***	−.09*	.75***	.65***	.60***	.89***	.63***	.53***	1	.72***	.64***	−.27***	.01	−.11*
10. Actual Grade T2	−.26***	−.11**	.60***	.74***	.68***	.63***	.84***	.67***	.65***	1	.67***	−.10**	−.19***	−.04
11. Actual Grade T3	−.34***	−.15**	.59***	.69***	.76***	.60***	.73***	.85***	.65***	.73***	1	−.05	−.03	−.33***
12. Self-Enhancement T1	.03	.05	.01	−.04	.01	.26***	.03	.05	−.21***	−.01	−.05	1	.10*	.08
13. Self-Enhancement T2	.02	.01	.06	.08	.09	.07	.34***	−.02	.03	−.23***	.03	.11***	1	.07
14. Self-Enhancement T3	−.06	−.03	−.03	−.05	.01	−.08	−.08	.31**	−.16***	−.03	−.25***	.14**	−.09	1

Note. ACSC = Academic self-concept; S-R = Self-reported; T1 = Time 1 assessment (N = 916); T2 = Time 2 assessment (N = 719); T3 = Time 3 assessment (N = 647).
$^{*}p \leq .05$. $^{**}p \leq .01$. $^{***}p \leq .001$.

displayed in Figure 1, and the correlations among the latent growth parameters represent residual correlations (i.e., controlled for the gender, age, and academic subject covariates). Regarding the associations among intercepts (i.e., initial scores), the intercept of self-enhancement was found to be positively associated with the intercept of self-concept, but not with the intercept of achievement. Moreover, the intercept of self-concept was associated with the intercept of achievement. Thus, higher initial scores of self-enhancement were associated with higher initial scores of self-concept, which were in turn associated with higher initial scores of achievement.

Longitudinal Interplay Between Self-Enhancement, Academic Self-Concept, and Academic Achievement

Associations among slopes (i.e., linear change over time). The slope of self-enhancement was found to be significantly associated neither with the slope of self-concept, nor with the slope of achievement. However, the slope of self-concept was positively associated with the slope of achievement. Therefore, more positive slopes of self-concept were associated with more positive slopes of achievement.

Associations between intercepts and slopes. The intercept of self-enhancement was negatively associated with the slope of self-concept. Furthermore, the intercept of self-concept was found to be negatively associated with the slopes of both self-concept and achievement. Finally, the intercept of achievement was negatively associated with the slope of self-concept. In other words, students with higher initial scores of self-enhancement were found to have more negative slopes of self-concept. Students with higher initial scores of self-concept were found to have more negative slopes of both self-concept and achievement. Students with higher initial scores of achievement were found to have more negative slopes of self-concept. All other associations between intercepts and slopes were nonsignificant.

Regarding the effects of the covariates on the latent growth parameters of self-enhancement, self-concept, and achievement (see Table 5), we found that males had a slightly higher intercept of self-enhancement, as well as a lower intercept and more negative slope of both self-concept and achievement. Age was found to be positively associated with the intercept of self-enhancement. As for the effect of the academic subject where mathematics was the reference category,

Table 3

Model Fit Indices for the Three Univariate LGMs and for the Two TPPLGMs

Model	χ^2	*df*	*p*	CFI	RMSEA	SRMR
Univariate LGM for Self-Enhancement	.097	1	.755	1.000	.000	.003
Univariate LGM for Academic Self-Concept	2.454	1	.117	.999	.019	.010
Univariate LGM for Academic Achievement	2.347	1	.125	.998	.019	.015
TPPLGM without Covariates	23.534	9	.005	.997	.020	.019
TPPLGM with Covariates[a]	86.470	24	.001	.988	.027	.018

Note. LGM = latent growth model; TPPLGM = two-part latent growth model.
[a] Gender, age, and academic subject (with mathematics as the reference category) were included as covariates in this model.

German was found to have a higher intercept of self-concept, as well as a higher intercept and a more positive slope of achievement. English was found to have a higher intercept and a more negative slope of self-concept, as well as a higher intercept of achievement. Finally, French did not differ from mathematics on any of the latent growth parameters. It must be noted that these results were controlled for the effect of the respective other covariates and that effect sizes were found to be quite low.

Indirect effect of self-enhancement on academic achievement. Although there was no statistically significant association between the intercept of self-enhancement and the slope of achievement, the possibility of an exclusively indirect association between these two growth parameters could still be examined. To this end, an indirect effect was modeled within the TPPLGM to test whether the association between the initial score of self-enhancement and the slope of achievement could be explained by the initial score of self-concept. The rationale for the selection of direct paths to be modeled was based on two considerations. First, self-enhancement was discussed as a strategy to enhance one's self-concept in the short term. This was modeled as a direct path (as opposed to the correlation reported above) from the intercept of self-enhancement to the intercept of self-concept. Second, self-concept has been found to be associated with increases in achievement within the same academic subject (Marsh, 1986; Marsh & Craven, 2006; Möller et al., 2011; Niepel et al., 2014). This was modeled as a direct path from the intercept of self-concept to the slope of achievement. In addition, we tested if the indirect effect from the intercept of self-enhancement to the intercept of self-concept and on to the slope of achievement was significant. The resulting model was found to fit the data well, as it was equivalent to the TPPLGM with covariates. Results of the TPPLGM with the indirect effect suggested that the effect of the intercept of self-enhancement on the slope of achievement could be explained by the intercept of self-concept ($\beta = -.06$; $p < .001$). Students that displayed self-enhancement in Grade 9 showed a higher self-concept in Grade 9 ($\beta = .28$; $p < .001$), which in turn led to a more negative slope in their achievement from Grade 9 to Grade 11 ($\beta = -.28$; $p < .01$).

Discussion

The purpose of the present study was to examine the longitudinal interplay between self-enhancement, self-concept, and achievement among students progressing from Grades 9 to 11. First, cross-sectional results on the association between self-enhancement, self-concept, and achievement were replicated. Second, long-term effects of self-enhancement on self-concept and achievement were explored. Finally, indirect long-term effects of self-enhancement on achievement that could be partially explained by an inflated self-concept were examined. To fulfill these objectives, a trivariate parallel process latent growth model (TPPLGM) was employed.

Cross-Sectional Relations Between Self-Enhancement, Academic Self-Concept, and Academic Achievement

The cross-sectional portion of the analyses (i.e., from the TPPLGM) indicated that, in the short term (i.e., during the same school year), students with high scores of self-enhancement showed higher scores of self-concept (controlling for gender, age, and academic subject). This result confirms our hypothesis and is in line with previous research on the effect of self-enhancement on self-esteem (Robins & Beer, 2001) and self-concept (Dickhäuser & Plenter, 2005). Furthermore, no direct association was found between self-enhancement and achievement in the short term. This result also confirms our expectation and is consistent with previous research (Robins & Beer, 2001). As for the cross-sectional association between self-concept and achievement, the TPPLGM yielded a very high and positive correlation between these constructs, which is in line with our hypothesis and

Table 4

Correlations Between Latent Growth Parameters of the TPPLGM Without (Above the Diagonal) and With Covariates (Gender, Age, and Academic Subject; Below the Diagonal)

Latent growth parameter	*M*	*SD*	1	2	3	4	5	6
1. Intercept Self-Enhancement	.09***	.10**	1	−.31	.21*	−.28*	.02	−.30
2. Slope Self-Enhancement	.02**	.05	−.29	1	−.13	.28	−.23	.20
3. Intercept Self-Concept	3.25***	1.01***	.28*	−.16	1	−.53***	.82***	−.25**
4. Slope Self-Concept	−.15***	.38***	−.35*	.42	−.52***	1	−.32***	.83***
5. Intercept Achievement	4.58***	.54***	.08	−.23	.83***	−.35***	1	−.08
6. Slope Achievement	−.04**	.18**	−.29	.24	−.28***	.85***	−.13	1

Note. TPPLGM = trivariate parallel process latent growth model; *M* = Mean; *SD* = Standard deviation. The mean scores (*M*) and the standard deviations (*SD*) of the latent growth parameters refer to the TPPLGM without covariates.
* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

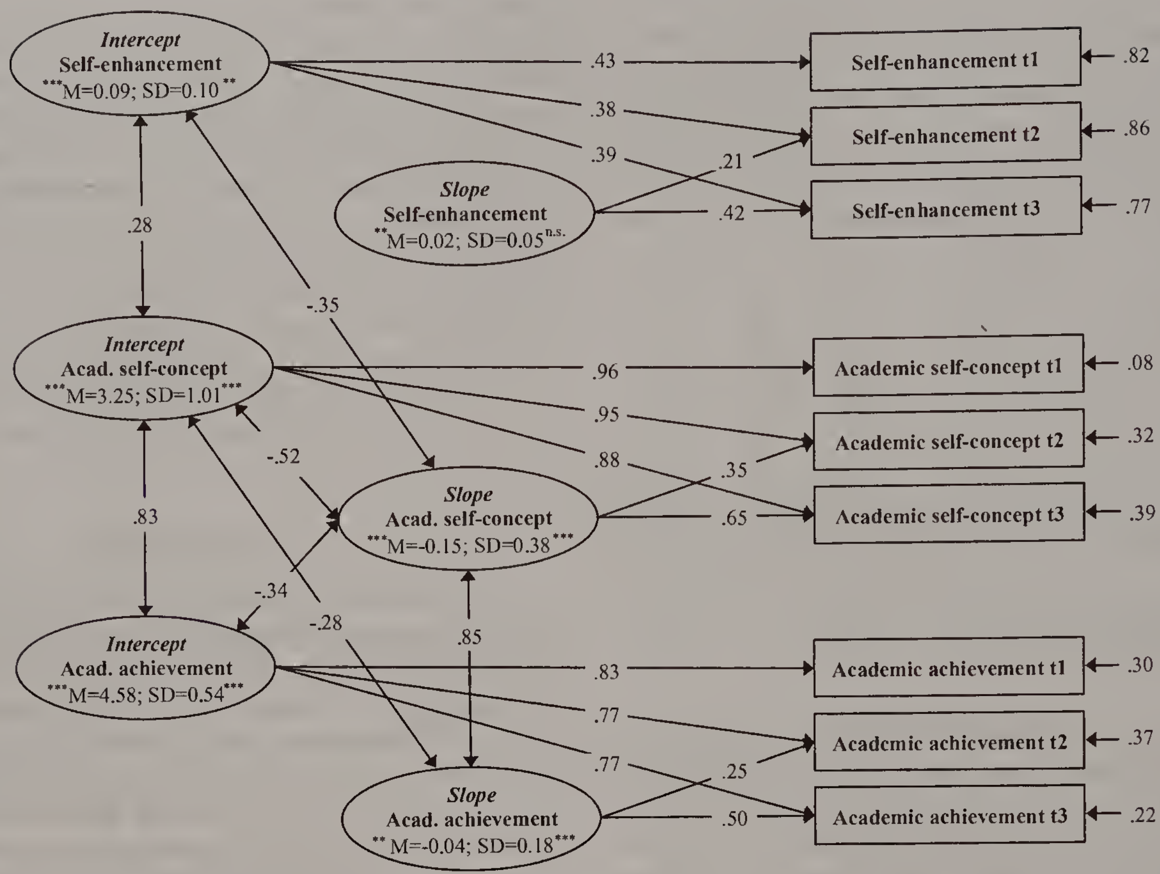


Figure 1. Standardized solution of the TPPLGM with covariates (gender, age, and academic subject). The mean scores (*M*) and the standard deviations (*SD*) of the latent growth parameters refer to the TPPLGM without covariates. The mean scores and the standard deviations were included here for a better overview. All correlations between latent growth parameters are indicated with straight double-headed arrows and represent *residual correlations* (i.e., correlations between the variance that was not explained by the covariates). Only significant correlations are displayed. *Note.* n.s. = not significant. $^{**}\ p \leq .01$. $^{***}\ p \leq .001$.

past findings (Marsh, 1986; Marsh & Craven, 2006; Niepel et al., 2014). In sum, self-enhancement appears to be an adaptive strategy in the short term—it was associated with a better self-concept, which was in turn positively associated with achievement. However, there was no direct association between self-enhancement and achievement in the short term.

Longitudinal Interplay Between Self-Enhancement, Academic Self-Concept, and Academic Achievement

Self-enhancement and academic self-concept. The longitudinal portion of the TPPLGM indicated that, in the long term (i.e., over multiple school years), self-enhancement was associated with a stronger decrease in self-concept. Thus, students with higher initial scores of self-enhancement tended to have more pronounced decreases in self-concept, which is in line with our hypothesis and with results obtained by Robins and Beer (2001) on the long-term effect of self-enhancement on self-esteem. Notably, the reverse association was not found to be significant—that is, the initial score of self-concept was not associated with the slope of self-enhancement. Although causality cannot be proven with these models, this pattern of associations supports the notion that self-enhancement drives changes in self-concept, not vice versa.

Academic self-concept and academic achievement. As for the relation between self-concept and achievement, our results suggest that their longitudinal association is reciprocal and negative. On the one hand, higher initial self-concept was associated with a stronger decrease in achievement. On the other hand, higher initial achievement was associated with a stronger decrease in self-concept. These results are in contrast to the hypotheses of the current study and initially appear to be inconsistent with the findings from previous research (Marsh, 1986; Marsh & Craven, 2006; Niepel et al., 2014). However, these results must be interpreted with caution and in the context of the existing associations among intercepts and slopes rather than in isolation. To this end, estimated growth trajectories of achievement for students with low (i.e., one *SD* below the mean), mean, and high (i.e., one *SD* above the mean) initial scores of self-concept were

Table 5
Standardized Regression Coefficients of the Effects of the Covariates on the Latent Growth Parameters

	Self-Enhancement		Self-Concept		Achievement	
Variable	Intercept	Slope	Intercept	Slope	Intercept	Slope
Male	.21*	−.04	−.06***	−.07*	−.17***	−.09*
Age	.14*	−.06	−.03	−.03	−.02	−.13
German	.03	−.14	.07*	.06	.12**	.23**
English	−.02	−.05	.13***	−.20***	.11**	.09
French	.03	−.18	.01	−.07	.02	.01

$^{*}\ p \leq .05$. $^{**}\ p \leq .01$. $^{***}\ p \leq .001$.

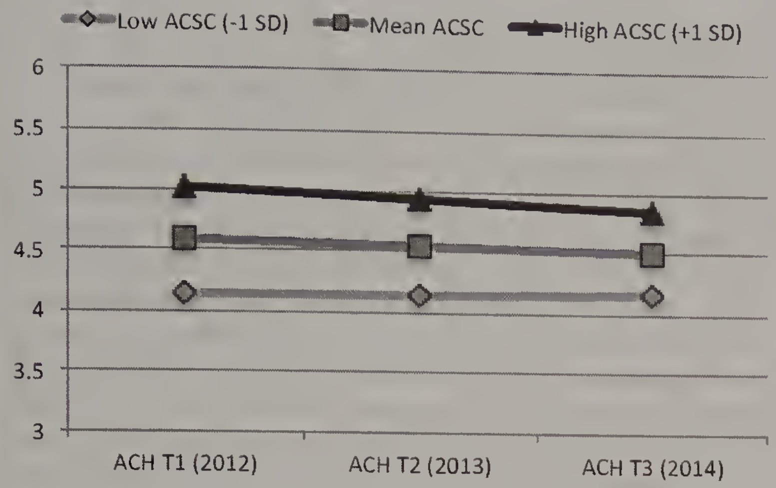


Figure 2. Estimated growth curves of academic achievement (ACH) for students with low ($M - 1$ SD), mean, and high ($M + 1$ SD) initial scores of academic self-concept (ACSC).

computed (see Figure 2). These additional results revealed that students who scored higher on self-concept at the first time point showed a more marked decrease in achievement over time. Importantly, these were also students with higher achievement scores at the first time point. Thus, although their decrease in achievement over time was more pronounced, these students tended to score highest on achievement relative to their peers, which is consistent with previous studies (e.g., Möller et al., 2011; Niepel et al., 2014). The same rationale applies to the association between the intercept of achievement and the growth of academic self-concept (see Figure 3).

Self-enhancement and academic achievement. Consistent with our expectation based on the results obtained by Robins and Beer (2001), we did not find any direct association between the initial level of self-enhancement and the development of achievement. However, our results suggest that the effect of self-enhancement on academic grades might be indirect rather than direct. High initial levels of self-enhancement were associated with higher levels of self-concept in the short term, and this inflated self-concept heightened the risk of a decrease in achievement over multiple years. In other words, our results suggest that the long-term decrease in achievement could be partly explained by the short-term increase in self-concept that is partly due to self-enhancement. In summary, self-enhancement was directly linked to a stronger decrease in self-concept and indirectly to a long-term decrease in achievement. These results extend the findings reported by Robins and Beer and, in so doing, do not support the notion that self-enhancement acts as a motivator in the face of adversity, thereby leading to better performance in the long term (Taylor & Brown, 1988, 1994).

Self-Enhancement as a Risk Factor for Declines in Academic Self-Concept and Achievement

On average, self-enhancement was found to slightly increase from Grades 9 to 11. Considering that self-enhancement seems to have long-term disadvantages, one might ask why some students keep self-enhancing. The answer may be that they do it because of the short-term advantages. Indeed, Robins and Beer (2001) acknowledged that self-enhancement can be a strategy by which students regulate their affect and self-esteem in situations that pose a threat to the self, which is particularly pronounced in individuals scoring high on narcissism. This strategy might work well in the short term, but, as it is based on unrealistic self-evaluation, it represents a risk for a decrease in self-concept and achievement in the long term (Robins & Beer, 2001). Our results support this notion, as self-enhancement artificially increased self-concept in the short term. This unrealistic increase makes the attainment of one's expectations equally unrealistic, which might result in lower achievement than expected. If expectations are not met, the self-concept is threatened, and self-enhancement might be triggered again to maintain a stable self-view. Thus, a vicious cycle could arise from these dynamics.

The role of learning effort. Learning effort is a variable that could explain why an inflated self-concept (as a result of self-enhancement) might lead to a stronger decrease in achievement in different academic subjects. Svanum and Bigatti (2006) found that uninformed and wishful optimism, which might also be interpreted as an indicator of self-enhancement, was associated with a lack of learning effort and learning skills (e.g., problem solving, critical thinking, metacognition), especially for students with low ability. Robins and Beer (2001) also discussed that learning effort might play a moderating role in this regard: If self-enhancement leads to an increase in one's self concept *and* is accompanied by greater learning efforts, achievement is likely to remain stable or to increase. If, however, self-enhancement leads to an increase in one's self-concept *and* is accompanied by lower effort or excessive procrastination, achievement may be negatively affected in the long term. Thus, it is possible that students that self-enhance tend to underestimate the effort that is needed to attain a certain level of achievement, which increases the likelihood that they invest an insufficient amount of effort. In contrast, Svanum and Bigatti found that students that displayed informed and aspirational optimism were more likely to invest appropriate effort and attain better grades. Accordingly, learning effort might moderate the indirect effect of self-enhancement on achievement. However, the potential moderating role of learning effort remains hypothetical and would need to be examined in future studies.

The role of causal attributions. Causal attributions (Weiner, Heckhausen, & Meyer, 1972) might play an important reinforcing role in the processes described above. There is some evidence indicating that high self-enhancement is accompanied by protective attributions in case of failure to meet one's expectations (Buckelew et al.,

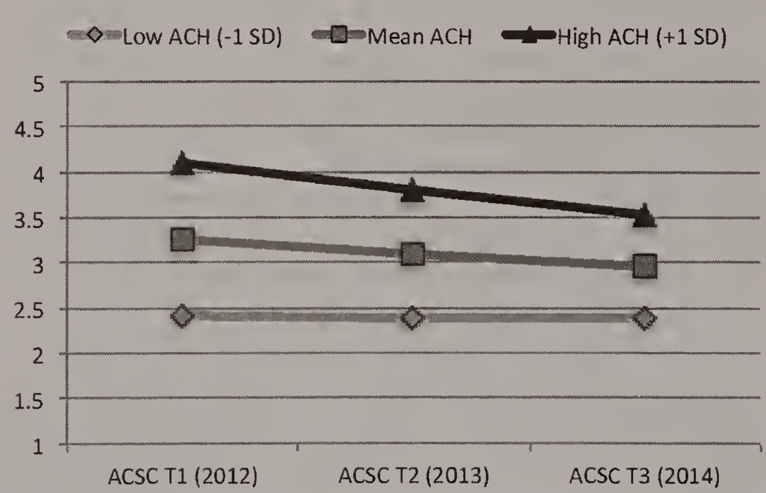


Figure 3. Estimated growth curves of academic self-concept (ACSC) for students with low ($M - 1$ SD), mean, and high ($M + 1$ SD) initial scores of academic achievement (ACH).

2013; Robins & Beer, 2001). In particular, it has been shown that self-enhancers tend to attribute success to internal and stable causes (e.g., ability) while at the same time attributing failure to external and unstable causes (e.g., luck). This self-protective strategy might prevent the student from realizing the reason(s) for his or her suboptimal grades, which in turn might put the student at risk for continued self-enhancement, an unrealistic self-concept, and the resulting decreases in achievement. Thus, the aforementioned vicious cycle may be reinforced by these attributions so that the likelihood of change in learning behavior is reduced (Buckelew et al., 2013). Over time, the strategy of self-enhancement will likely become less effective as the decrease in achievement will inevitably affect the self-concept causing it to decline. In the worst case, self-enhancers might choose to disengage from the academic context (Robins & Beer, 2001) and disregard its importance for the self, which could lead to a stronger decrease in achievement.

In sum, the effect of self-enhancement on achievement seems to be quite complex and involves a number of moderating (e.g., effort) and mediating (e.g., attributions) factors. Accordingly, future research might explore the longitudinal mediating and moderating roles of these variables. Such knowledge would enhance our understanding of the conditions under which different forms of self-enhancement influence self-concept and achievement in the long term. Subsequent efforts could be taken to design interventions aimed at optimizing learning strategies, effort, and achievement.

Over-Reporting and Over-Confident Calibration: Two Sides of the Same Coin?

Exaggerated self-evaluations have been found with respect to many personal characteristics and have been examined from different points of view and under different labels such as self-observer rating discrepancies (e.g., Nilsen & Campbell, 1993), optimisms (Weinstein, 1980), positive illusions and creative self-deception (Taylor & Brown, 1994), calibration (Alexander, 2013), and self-enhancement (Sedikides & Gregg, 2008). In the present study, we focused on overreporting of past grades as a form of self-enhancement that needs to be distinguished from other forms of self-enhancement that refer to future events. As outlined in the introduction, there is a seemingly small but quite important difference between overreporting past achievement and being overconfident (i.e., low calibration) about future achievement: certainty versus uncertainty. To date, no study has examined whether these two forms of self-enhancement are differentially linked to self-concept and achievement (or other constructs) in the long term. Results from the present study and from research on overestimating grades (Buckelew et al., 2013) and on calibration (Chiu & Klassen, 2010) suggest that all forms of self-enhancement might have undesirable long-term effects, including less learning effort and external attributions (see above). Future studies might therefore want to explore these differential long-term effects and/or shed light on the associations between forms of self-enhancement that refer to past and to future events. It might be that students that overreport past grades also tend to overestimate future grades because the underlying mechanism can be assumed to be the same, namely self-enhancement.

Implications for Practice

The results of the present study suggest that greater self-enhancement can result in a stronger decrease in self-concept and achievement in the long term. It is important to note that there is not a threshold above which self-enhancement turns from a *positive* effect to a *negative* long-term effect on self-concept and achievement. Yet self-enhancement is linearly and negatively associated with self-concept and achievement such that greater initial self-enhancement equals greater decline in self-concept and achievement over time. In light of these results, one implication for practice is that it might be crucial to educate students on the importance of accurate self-evaluations and realistic self-expectations regarding both past and future achievements (i.e., accurate calibration of self-concept). Furthermore, as Buckelew et al. (2013) discussed, it is important to increase students' awareness of the potentially negative effects of external attributions following failure and to train them to develop an attributional style that leads to higher school engagement and adaptive coping strategies. As Svanum and Bigatti (2006) acknowledged, this does not mean that optimism needs to be curbed or discouraged. Rather, students need clarification and instruction regarding the skills and commitment needed to attain their expected level of achievement. This might be particularly important for students with comparably low ability, as previous research has shown that the tendency to self-enhance is notably pronounced among this group, possibly resulting from perceived added pressure to obtain better grades (Minkov, 2008; Schwartz & Beaver, 2015). Unfortunately, they may tend to use self-enhancement to regulate their self-concept in the short term but are unable to follow up with appropriate learning efforts, which might negatively affect their achievement.

Strengths and Limitations

To the best of our knowledge, this is the first study to adopt a *trivariate* longitudinal approach to examine the interplay between the latent developments of self-enhancement, self-concept, and achievement. This approach yielded the first empirical findings on the reciprocal longitudinal relations between these three constructs and provided initial insight into the complexity of their relations. The latent nature of the growth models used to address the present research questions reinforced the validity of our results, as all models that were computed showed a very good fit to the data. Moreover, we used an objective criterion to assess self-enhancement, namely students' actual academic grades. Furthermore, data were collected on four academic subjects, which reinforces the generalizability of the present results. Additionally, gender and age were controlled for in all of our models.

The present study is not without limitations. First, operationalizing self-enhancement as the difference between self-reported and actual grades has the drawback that students with the best possible grade could not overreport their grade. However, averaged across all school subjects and measurement occasions, only 2.6% of the students had the highest grade (median 2.7%, max. 5.1%, min. 0.2%). Thus this limitation likely had little influence on the obtained results. Second, it must also be noted that the time frame under consideration was limited to two years, and the sample was composed solely of high-school students. Future studies might examine the effects of self-enhancement on self-concept and achievement across an extended time period and among primary school and university students. Third, despite being relatively large, the present sample was not representa-

tive of the Swiss population of ninth- to eleventh-graders, as the French- and Italian-speaking populations of Switzerland were not represented. Additional studies are needed to assess the external validity of the results reported herein. Furthermore, gender, age, and academic subject were controlled for in the present analyses while other potentially relevant covariates were excluded such as parental education and income (Shaw & Mattern, 2009), time since graduation and ethnicity (Talento-Miller & Peyton, 2006), and genetic and environmental influences (Schwartz & Beaver, 2015). Fourth, the ICC of academic achievement was found to be relatively low (i.e., .05). This result might be due to high homogeneity between high-school classes in our sample or because grading on the curve practices lead to reduced variance between classes. Finally, we did not explore associations across different academic subjects, which would be an important next step as previous findings suggest that associations within an academic subject are different from those across academic subjects (Möller et al., 2011; Niepel et al., 2014).

Conclusion

The present study yielded the first results showing that self-enhancement is associated with a higher academic self-concept in the short term and that this short-term increase might lead to a stronger decrease in achievement over time. An inflated self-concept might lead to unrealistic expectations and less efficient learning strategies or reduced learning efforts, which in turn may lead to lower achievement. If a decrease in achievement is then attributed to external causes, the likelihood of continued self-enhancement increases and a vicious cycle may arise and lead to a decrease in achievement in the long term. Thus, the reciprocal associations between self-enhancement, self-concept, and achievement are highly complex and involve a number of mediating (e.g., attributions) and moderating (e.g., learning effort) variables that need to be examined in more detail.

References

Ackerman, P. L., & Wolman, S. D. (2007). Determinants and validity of self-estimates of abilities and self-concept measures. *Journal of Experimental Psychology: Applied, 13,* 57–78. http://dx.doi.org/10.1037/1076-898X.13.2.57

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction, 24,* 1–3. http://dx.doi.org/10.1016/j.learninstruc.2012.10.003

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49,* 1621–1630. http://dx.doi.org/10.1037/0022-3514.49.6.1621

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73,* 269–290. http://dx.doi.org/10.3200/JEXE.73.4.269-290

Bollen, K. A., & Curran, P. J. (2005). *Latent curve models: A structural equation perspective* (1st ed.). Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/0471746096

Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgments. *Social Cognition, 4,* 353–376. http://dx.doi.org/10.1521/soco.1986.4.4.353

Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin, 38,* 209–219. http://dx.doi.org/10.1177/0146167211432763

Buckelew, S. P., Byrd, N., Key, C. W., Thornton, J., & Merwin, M. M. (2013). Illusions of a good grade: Effort or luck? *Teaching of Psychology, 40,* 134–138. http://dx.doi.org/10.1177/0098628312475034

Chiu, M. M., & Klassen, R. M. (2010). Relations of mathematics self-concept and its calibration with mathematics achievement: Cultural differences among fifteen-year-olds in 34 countries. *Learning and Instruction, 20,* 2–17. http://dx.doi.org/10.1016/j.learninstruc.2008.11.002

Chung, T., White, H. R., Hipwell, A. E., Stepp, S. D., & Loeber, R. (2010). A parallel process model of the development of positive smoking expectancies and smoking behavior during early adolescence in Caucasian and African American girls. *Addictive Behaviors, 35,* 647–650. http://dx.doi.org/10.1016/j.addbeh.2010.02.005

Crocker, J., Karpinski, A., Quinn, D. M., & Chase, S. K. (2003). When grades determine self-worth: Consequences of contingent self-worth for male and female engineering and psychology majors. *Journal of Personality and Social Psychology, 85,* 507–516. http://dx.doi.org/10.1037/0022-3514.85.3.507

Dickhäuser, O., & Plenter, I. (2005). Letztes Halbjahr stand ich zwei [On the accuracy of self-reported school marks]. *Zeitschrift fur Padagogische Psychologie, 19,* 219–224. http://dx.doi.org/10.1024/1010-0652.19.4.219

Goetz, T., Ehret, C., Jullien, S., & Hall, N. C. (2006). Is the grass always greener on the other side? Social comparisons of subjective well-being. *The Journal of Positive Psychology, 1,* 173–186. http://dx.doi.org/10.1080/17439760600885655

Goetz, T., Sticca, F., Pekrun, R., Murayama, K., & Elliot, A. J. (2016). Intraindividual relations between achievement goals and discrete achievement emotions: An experience sampling approach. *Learning and Instruction, 41,* 115–125. http://dx.doi.org/10.1016/j.learninstruc.2015.10.007

Gramzow, R. H., & Willard, G. (2006). Exaggerating current and past performance: Motivated self-enhancement versus reconstructive memory. *Personality and Social Psychology Bulletin, 32,* 1114–1125. http://dx.doi.org/10.1177/0146167206288600

Heinrich, C. J., & Lynn, L. E. (2001). Means and ends: A comparative study of empirical methods for investigating governance and performance. *Journal of Public Administration: Research and Theory, 11,* 109–138. http://dx.doi.org/10.1093/oxfordjournals.jpart.a003490

Huang, F. L. (2016). Alternatives to multilevel modeling for the analysis of clustered data. *Journal of Experimental Education, 84,* 175–196. http://dx.doi.org/10.1080/00220973.2014.952397

King, K. M., Nguyen, H. V., Kosterman, R., Bailey, J. A., & Hawkins, J. D. (2012). Co-occurrence of sexual risk behaviors and substance use across emerging adulthood: Evidence for state- and trait-level associations. *Addiction, 107,* 1288–1296. http://dx.doi.org/10.1111/j.1360-0443.2012.03792.x

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75,* 63–82. http://dx.doi.org/10.3102/00346543075001063

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35,* 125–141. http://dx.doi.org/10.1207/S15326985EP3502_6

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23,* 129–149. http://dx.doi.org/10.3102/00028312023001129

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1,* 133–163. http://dx.doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III: The construct validity of multidimensional self-concept ratings by late

adolescents. *Journal of Educational Measurement, 21,* 153–174. http://dx.doi.org/10.1111/j.1745-3984.1984.tb00227.x

McCrea, S. M., & Hirt, E. R. (2001). The role of ability judgments in self-handicapping. *Personality and Social Psychology Bulletin, 27,* 1378–1389. http://dx.doi.org/10.1177/01461672012710013

Minkov, M. (2008). Self-enhancement and self-stability predict school achievement at the national level. *Cross-Cultural Research: The Journal of Comparative Social Science, 42,* 172–196. http://dx.doi.org/10.1177/1069397107312956

Möller, J., Retelsdorf, J., Köller, O., & Marsh, H. W. (2011). The reciprocal internal/external frame of reference model: An integration of models of relations between academic achievement and self-concept. *American Educational Research Journal, 48,* 1315–1346. http://dx.doi.org/10.3102/0002831211419649

Möller, J., Streblow, L., Pohlmann, B., & Köller, O. (2006). An extension to the internal/external frame of reference model to two verbal and numerical domains. *European Journal of Psychology of Education, 21,* 467–487. http://dx.doi.org/10.1007/BF03173515

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Niepel, C., Brunner, M., & Preckel, F. (2014). The longitudinal interplay of students' academic self-concepts and achievements within and across domains: Replicating and extending the reciprocal internal/external frame of reference model. *Journal of Educational Psychology, 106,* 1170–1191. http://dx.doi.org/10.1037/a0036307

Nilsen, D., & Campbell, D. P. (1993). Self–observer rating discrepancies: Once an overrater, always an overrater? *Human Resource Management, 32,* 265–281. http://dx.doi.org/10.1002/hrm.3930320206

Noble, R. N., Heath, N. L., & Toste, J. R. (2011). Positive illusions in adolescents: The relationship between academic self-enhancement and depressive symptomatology. *Child Psychiatry and Human Development, 42,* 650–665. http://dx.doi.org/10.1007/s10578-011-0242-5

Pekrun, R., Hall, N. C., Goetz, T., & Perry, R. P. (2014). Boredom and academic achievement: Testing a model of reciprocal causation. *Journal of Educational Psychology, 106,* 696–710. http://dx.doi.org/10.1037/a0036006

Preckel, F., Niepel, C., Schneider, M., & Brunner, M. (2013). Self-concept in adolescence: A longitudinal study on reciprocal effects of self-perceptions in academic and social domains. *Journal of Adolescence, 36,* 1165–1175. http://dx.doi.org/10.1016/j.adolescence.2013.09.001

Rhodewalt, F., Morf, C., Hazlett, S., & Fairfield, M. (1991). Self-handicapping: The role of discounting and augmentation in the preservation of self-esteem. *Journal of Personality and Social Psychology, 61,* 122–131. http://dx.doi.org/10.1037/0022-3514.61.1.122

Robins, R. W., & Beer, J. S. (2001). Positive illusions about the self: Short-term benefits and long-term costs. *Journal of Personality and Social Psychology, 80,* 340–352. http://dx.doi.org/10.1037/0022-3514.80.2.340

Ross, M., & Wilson, A. E. (2003). Autobiographical memory and conceptions of self getting better all the time. *Current Directions in Psychological Science, 12,* 66–69. http://dx.doi.org/10.1111/1467-8721.01228

Rusbult, C. E., Van Lange, P. A., Wildschut, T., Yovetich, N. A., & Verette, J. (2000). Perceived superiority in close relationships: Why it exists and persists. *Journal of Personality and Social Psychology, 79,* 521–545. http://dx.doi.org/10.1037/0022-3514.79.4.521

Schneider, R., & Sparfeldt, J. R. (2016). Zur (Un-)Genauigkeit selbstberichteter Zensuren bei Grundschulkindern [The accuracy of self-reported grades in elementary school]. *Psychologie in Erziehung und Unterricht, 63,* 48–59. http://dx.doi.org/10.2378/peu2016.art05d

Schwartz, J. A., & Beaver, K. M. (2015). Making (up) the grade? Estimating the genetic and environmental influences of discrepancies between self-reported grades and official GPA scores. *Journal of Youth and Adolescence, 44,* 1125–1138. http://dx.doi.org/10.1007/s10964-014-0185-9

Schwinger, M., Wirthwein, L., Lemmer, G., & Steinmayr, R. (2014). Academic self-handicapping and achievement: A meta-analysis. *Journal of Educational Psychology, 106,* 744–761. http://dx.doi.org/10.1037/a0035832

Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science, 3,* 102–116. http://dx.doi.org/10.1111/j.1745-6916.2008.00068.x

Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. *Advances in Experimental Social Psychology, 29,* 209–269. http://dx.doi.org/10.1016/S0065-2601(08)60018-0

Shaw, E. J., & Mattern, K. D. (2009). *Examining the accuracy of self-reported high school grade point average* (College Board Research Rep. No. 9–5). New York, NY: College Board.

Sparfeldt, J. R., Buch, S. R., Rost, D. H., & Lehmann, G. (2008). Akkuratesse selbstberichteter Zensuren [The accuracy of self-reported grades in school]. *Psychologie in Erziehung und Unterricht, 55,* 68–75.

Stone, C. A., & May, A. L. (2002). The accuracy of academic self-evaluations in adolescents with learning disabilities. *Journal of Learning Disabilities, 35,* 370–383. http://dx.doi.org/10.1177/00222194020350040801

Svanum, S., & Bigatti, S. (2006). Grade expectations: Informed or uninformed optimism, or both? *Teaching of Psychology, 33,* 14–18. http://dx.doi.org/10.1207/s15328023top3301_4

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica, 47,* 143–148. http://dx.doi.org/10.1016/0001-6918(81)90005-6

Talento-Miller, E., & Peyton, J. (2006). *Moderators of the accuracy of self-report Grade Point Average* (GMAC Research Rep. No. 06–10). McLean, VA: Graduate Management Admission Council.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103,* 193–210. http://dx.doi.org/10.1037/0033-2909.103.2.193

Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin, 116,* 21–27. http://dx.doi.org/10.1037/0033-2909.116.1.21

Vancouver, J. B., & Kendall, L. N. (2006). When self-efficacy negatively relates to motivation and performance in a learning context. *Journal of Applied Psychology, 91,* 1146–1153. http://dx.doi.org/10.1037/0021-9010.91.5.1146

Weiner, B., Heckhausen, H., & Meyer, W. U. (1972). Causal ascriptions and achievement behavior: A conceptual analysis of effort and reanalysis of locus of control. *Journal of Personality and Social Psychology, 21,* 239–248. http://dx.doi.org/10.1037/h0032307

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39,* 806–820. http://dx.doi.org/10.1037/0022-3514.39.5.806

Willard, G., & Gramzow, R. H. (2008). Exaggeration in memory: Systematic distortion of self-evaluative information under reduced accessibility. *Journal of Experimental Social Psychology, 44,* 246–259. http://dx.doi.org/10.1016/j.jesp.2007.04.012

Wojcik, S. P., & Ditto, P. H. (2014). Motivated happiness self-enhancement inflates self-reported subjective well-being. *Social Psychological and Personality Science, 5,* 825–834. http://dx.doi.org/10.1177/1948550614534699

Received December 14, 2015
Revision received October 5, 2016
Accepted November 17, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 6, 855–870

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000175

Fish Swimming Into the Ocean: How Tracking Relates to Students' Self-Beliefs and School Disengagement at the End of Schooling

Hanna Dumont
German Institute for International Educational Research, Berlin, Germany

Paula Protsch
WZB Berlin Social Science Center, Berlin, Germany

Malte Jansen
German Institute for International Educational Research, Berlin, Germany

Michael Becker
German Institute for International Educational Research, Berlin, Germany, and Leibniz Institute for Science and Mathematics Education, Kiel, Germany

In this study, we analyzed how secondary school tracking relates to students' self-beliefs (i.e., their academic self-concepts in different domains and their beliefs regarding their labor market chances) and school disengagement during a time period that has received little attention in educational psychological research on tracking: when students are at the end of schooling and on the verge of entering the labor market. In doing so, we disentangled 2 distinguishing features of tracking: tracks as social contexts (operationalized via track level and the mean achievement of students' schoolmates) and tracks as pathways to different future opportunities (operationalized via educational certificates). Using questionnaire, achievement, and administrative school data from 2,155 students from 29 low-track schools, 23 intermediate-track schools, and 35 comprehensive schools in Berlin, Germany, we found educational certificates to be the most important factor shaping students' self-beliefs and school disengagement. Irrespective of their individual achievement, their schoolmates' achievement, and their track level, students who received the intermediate school-leaving certificate had higher academic self-concepts, believed that their certificate would give them better chances of success in the labor market, and were less disengaged from school than students who received the low school-leaving certificate. In contrast, students' track level did not serve as a predictor for the outcomes considered. The achievement of students' schoolmates (i.e., the big-fish-little-pond effect) was only relevant for students' academic self-concepts and not for students' self-beliefs regarding labor market entry or their school disengagement.

Keywords: tracking, academic self-concept, BFLPE, self-beliefs, educational certificates

All public school systems are faced with the challenge of how to efficiently organize learning processes while at the same time responding to each student's needs. Tracking—the grouping of students with similar achievement levels into different schools, study programs, or courses—is a common response to this challenge, in particular in secondary schooling. As a result, students develop and are socialized in different educational contexts (Pallas, Entwisle, Alexander, & Stluka, 1994). Because these contexts may provide different opportunities for learning and attainment, they may consequently contribute to educational inequality (Gamoran, 1992); hence, tracking practices have been heatedly debated in both policy and research. Indeed, studies have shown that track assignment is biased by the social backgrounds of students (Lucas & Berends, 2002; Maaz, Trautwein, Lüdtke, & Baumert,

This article was published Online First February 13, 2017.

Hanna Dumont, Department of Educational Governance, German Institute for International Educational Research, Berlin, Germany; Paula Protsch, Research Unit "Skill Formation and Labor Markets", WZB Berlin Social Science Center, Berlin, Germany; Malte Jansen, Department of Educational Governance, German Institute for International Educational Research; Michael Becker, Department of Educational Governance, German Institute for International Educational Research, Berlin, Germany, and Department of Educational Research, Leibniz Institute for Science and Mathematics Education, Kiel, Germany.

Malte Jansen is now at the Research Data Centre at the Institute for Educational Quality Improvement, Berlin, Germany.

This paper was partially funded by the College for Interdisciplinary Educational Research (a joint initiative of the BMBF, the Jacobs Foundation, and the Leibniz Association). We are grateful to the team of the BERLIN-study at the German Institute for International Educational Research, the Max-Planck Institute for Human Development, and the Leibniz Institute for Science and Mathematics Education for allowing us to use the dataset. The BERLIN-study is funded by the Berlin Senate Administration for Education, Science and Research and the Jacobs Foundation. We would like to thank Anna Katyn Chmielewski, Heike Solga, the participants of the colloquium at the WZB Berlin Social Science Center, and the anonymous reviewers for their very useful feedback on previous drafts. We would also like to thank Roisin Cronin for editorial assistance.

Correspondence concerning this article should be addressed to Hanna Dumont, Department of Educational Governance, German Institute for International Educational Research, Warschauer Str. 34-38, 10243 Berlin, Germany. E-mail: dumont@dipf.de

2008) and that students in lower tracks often learn less and have fewer postsecondary opportunities (Becker, Lüdtke, Trautwein, Köller, & Baumert, 2012; Brunello & Checchi, 2007; Oakes, 1985). From an educational psychological perspective, however, scholars have argued that being placed in a low track does not just have detrimental effects; in fact, when outcomes other than academic achievement or attainment are considered—namely students' academic self-concepts—the effects may be positive (Liu, Wang, & Parkins, 2005; Schwarzer, Lange, & Jerusalem, 1982; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006; Trautwein, Lüdtke, Marsh, & Nagy, 2009). This argument is based on the *big-fish-little-pond effect* (BFLPE), which posits that students' academic self-concepts are not just influenced by their individual achievement but also by their peers' achievement levels as a result of social comparison processes (Marsh, 1987). Accordingly, students feel more competent when they are surrounded by low-achieving peers, as is the case in lower tracks. This has led researchers to conclude that it is more beneficial for students' academic self-concepts to be a "big fish in a small pond" than to be a "small fish in a big pond" (Marsh, 1987; Marsh & Hau, 2003; Seaton, Marsh, & Craven, 2009). Educational psychological research on the BFLPE has thus focused on the effects of tracks as *social contexts*. In the present paper, we argue that this perspective should be broadened. As sociologists have pointed out (Gamoran, 1986; Lucas, 1999; Pallas et al., 1994), tracks not only constitute different immediate social contexts, but also provide students with different educational credentials that may be related to diverging *future opportunities*. This feature of tracking is particularly evident at the end of schooling, when students are on the verge of entering the labor market and are thus being exposed to the world beyond school. Whereas this time point in students' academic careers has been extensively studied by sociologists and developmental psychologists (e.g., Allmendinger, 1989; Brzinsky-Fay, 2007; M. Buchmann, & Kriesi, 2011; Heckhausen, Chang, Greenberger, & Chen, 2013; Protsch & Solga, 2015; Schoon, McCulloch, Joshi, Wiggins, & Bynner, 2001), it has not been a prominent topic in educational psychology. Moreover, because they have mainly studied the effects of tracking on academic self-concepts, educational psychologists have focused on quite specific self-beliefs. Self-beliefs generally refer to a person's beliefs about his or her attributes and abilities (Valentine, DuBois, & Cooper, 2004); by contrast, academic self-concepts are a more specific kind of self-belief and refer to a person's beliefs about his or her abilities in a particular academic domain and are commonly studied at the level of school subjects such as mathematics, English, or science (Marsh, 1990). However, there may well be other self-beliefs worth looking at when investigating the effects of tracking, especially when studying how students feel when they are about to finish school.

In the present paper, we extend previous research in educational psychology on tracking by bringing the sociological perspective into view and analyzing students' self-beliefs at the end of schooling. Speaking in terms of the BFLPE metaphor, we analyze what happens to the fish when they have to leave their pond and swim into the ocean. We do so within the context of the German educational system, which is a prime example of a rigid tracking system (Bol & Van de Werfhorst, 2013), and is thus particularly suited to studying tracking effects. In Germany, students are sorted into schools of different tracks right after elementary school. These school tracks can be regarded as different social contexts in which students learn and are socialized. In addition, school tracks often lead to different school-leaving certificates, which are related to different further educational and occupational pathways. Hence, in the German education system, students in different school tracks encounter different social contexts and different future opportunities. But although these two features of tracking used to be quite strongly tied together, reforms have been introduced in recent years to increase the permeability of the education system. These reforms have made it increasingly possible for students to receive different school-leaving certificates in different school tracks. In other words, students belonging to the same school track can receive different school-leaving certificates and students with the same school-leaving certificate may have experienced very different social contexts because they attended schools of different tracks. In our study, which uses data from the state of Berlin, we make use of this unique characteristic of the German tracking system in order to disentangle these two features of tracking and their effects on students' self-beliefs. In doing so, we not only focus on students' academic self-concepts, but also consider students' self-beliefs regarding their perceived future chances, namely students' self-beliefs about their labor-market entry opportunities. As a student's perception of limited future opportunities might lead him or her to disengage from school as a self-protective mechanism, we additionally investigate students' school disengagement to gain a more complete picture of the effects of tracking.

Two Distinguishing Features of Tracking

The practice of tracking can be found in almost all school systems around the world, mostly at the secondary level.[1] Even though the nature and extent of tracking varies greatly between countries, states, and/or school districts, all types of tracking have two distinguishing features. First, tracking creates distinct social contexts for students. The degree to which this is the case is mainly the result of the organizational level of tracking (Trautwein et al., 2006). That is, school systems can track students either between or within schools. In the former type of system, students of different achievement levels go to completely different schools, which often differ greatly in curricula. In the latter, all students go to the same school but are grouped together full-time for all subjects or part-time for some subjects, which allows students to take different course levels in different subjects. In some countries, there are also combinations of between- and within-school tracking. Taken together, the organizational level of tracking substantially determines whom students interact with on an everyday basis and thus the social context for students.

Second, tracking has an impact on students' future occupational and academic careers (Trautwein et al., 2006). This is particularly the case in countries such as Germany, where tracks typically lead to different educational certificates that substantially influence

[1] Some authors have also used the term *tracking* for the implicit grouping of students into schools by social background due to factors such as area of residence, namely *implicit school-level tracking* (Trautwein et al., 2006). In the present paper, we do not use this terminology and only use the term tracking for the deliberate sorting of students into different groups according to their achievement.

students' future occupational and educational paths. But even in countries where tracking is less salient and "visible", tracks are often associated with students' future opportunities. For instance, in the United States, high schools offer college preparatory or advanced placement courses, which can be thought of as high tracks and which influence students' chances of getting accepted at a good university.

Previous educational psychological research on the effects of tracking on students' self-beliefs has focused on how tracks function as social contexts and hence shape students' self-beliefs (Alicke, Zell, & Bloom, 2010; Chmielewski, Dumont, & Trautwein, 2013; Huguet et al., 2009; Marsh, 1990; Marsh, Chessor, Craven, & Roche, 1995; Thijs, Verkuyten, & Helmond, 2010; Trautwein et al., 2006). We summarize the findings from this research strand in the following section. Drawing on sociological research, we then turn to the second feature of tracking and discuss how the different future opportunities that tracks provide may also affect students' self-beliefs.

How Tracks Influence Students' Self-Beliefs as Social Contexts

Most of the research on how tracks influence students' self-beliefs as social contexts has focused on students' academic self-concepts. It is well established that a student's academic self-concept is shaped not only by his or her performance but also by social comparisons (Marsh et al., 1995). Two important social comparison mechanisms through which academic self-concept is known to be affected in tracking contexts are contrast and assimilation effects (Marsh et al., 1995; Marsh, Kong, & Hau, 2000). The *contrast effect* refers to the finding that students compare their own achievement with that of their class- or schoolmates, which leads them to feel more negative about their own competencies in a high-achieving group than in a low-achieving group (Marsh et al., 1995, 2000). This social comparison mechanism lies at the heart of the BFLPE (Marsh, 1987), which has been the subject of a great number of studies in the past 40 years (e.g., Bassis, 1977; Marsh, Hau, & Craven, 2004; Marsh & O'Mara, 2008; Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007; Nagengast & Marsh, 2012; Schwarzer et al., 1982; Seaton, Marsh, & Craven, 2010; Tymms, 2001; Zeidner & Schleyer, 1999). Empirically, the BFLPE is evident when there is a negative association between a group's mean achievement (usually on the school or class level) and a student's academic self-concept after controlling for the student's individual achievement; this has been replicated numerous times across many different educational systems (Marsh & Hau, 2003; Seaton et al., 2009). With respect to tracking, the BFLPE implies that a student's academic self-concept will benefit when in a lower track, because he or she is surrounded by students with low competencies and thus has fewer opportunities for upward comparisons. The positive consequences of tracking on academic self-concepts for students assigned to a low track have been shown in a large number of studies (Liem, Marsh, Martin, McInerney, & Yeung, 2013; Liu et al., 2005; Mulkey, Catsambis, Steelman, & Crain, 2005; Reuman, 1989; Schwarzer et al., 1982; Sung, Huang, Tseng, & Chang, 2014; Trautwein et al., 2006; Wouters, De Fraine, Colpin, Van Damme, & Verschueren, 2012).

The second social comparison mechanism that affects students' academic self-concept in tracking contexts is the *assimilation effect*, also known as the *basking in reflected glory* or *labeling effect* (Cialdini et al., 1976; Marsh et al., 1995, 2000). It is based on the assumption that tracks can be viewed as institutionalized educational categories that convey information to society at large about students' competencies (Pallas et al., 1994). Consequently, it states that being a member of a high track can make students feel positive about their own competencies, because they identify with the high track as a highly valued social group. Similarly, students in lower tracks may feel bad about their own competencies because it implies membership of a group with low prestige. In fact, some research has even proposed that students in lower tracks feel stigmatized (Solga, 2004). The assimilation effect should thus affect a student's academic self-concept in the opposite direction to the contrast effect. There are studies that have found the academic self-concepts of high-track students to be higher than those of low-track students (Chiu et al., 2008; Oakes, 1985), which can be seen as an empirical indication of assimilation effects. However, contrast and assimilation effects are not easy to study, as both effects occur at the same time.[2] There are some studies that have sought to disentangle both counterbalancing effects by simultaneously investigating how students' academic self-concept is affected by their track's mean achievement (as an operationalization of contrast effects) and their track membership (as an operationalization of assimilation effects) while controlling for students' individual achievement, but they have found mixed results (Marsh et al., 2000; Preckel & Brüll, 2010; Trautwein et al., 2006; Trautwein et al., 2009). A recent internationally comparative study by Chmielewski et al. (2013) suggested that the relative strength of these two counterbalancing social comparison mechanisms depends on the organizational level of tracking, as this determines whom students compare themselves to. With the exception of tracking systems in which students were grouped only for certain subjects, contrast effects outweighed assimilation effects, showing that, in most countries, students do indeed benefit from being in a lower track with respect to their academic self-concepts.

However, there are also a few studies that have investigated the consequences of tracking for other types of self-beliefs, and these have found disadvantages for low-track students and advantages for high-track students. For instance, Fuligni, Eccles, and Barber (1995) found that tracking in mathematics had a positive impact on intermediate- and high-track students' career-related self-concepts (as well as their future educational expectations), even though they did not find any differences in academic self-concept between students in different tracks. Van Houtte, Demanet, and Stevens (2012) showed that students in high tracks had higher self-esteem than students in vocational tracks, with the differences being more pronounced in within-school tracking than in between-school tracking systems. In order to explicitly test potential labeling or stigmatization processes associated with being in a low-track school, Knigge and Hannover (2011) investigated students' "collective identity" in different school tracks in Germany. They found that low-track school students had a negative collective identity and students at high-track schools had a very positive one. That is,

[2] As contrast and assimilation effects occur at the same time, the BFLPE should actually be regarded as the net effect of these two counterbalancing processes, with contrast effects outweighing assimilation effects (Huguet et al., 2009; Marsh, Kong, & Hau, 2000).

students in low-track schools had a more negative perception of what other people thought of their achievements, their motivation, and their social competence. Interestingly, this negative collective identity was accompanied by low school-related motivation. Along similar lines, some researchers have argued that students who believe they cannot succeed in school will disengage and reduce their efforts (Carbonaro, 2005; Kelly & Carbonaro, 2012). Indeed, Van Houtte and Stevens (2009) found that vocational track students had a weaker sense of school belonging than academic track students. The authors argued that students in lower tracks distance themselves from school in order to deal with their low social status.

How Tracks May Influence Self-Beliefs Through Diverging Future Opportunities: The Importance of Educational Credentials

In addition to being distinct social contexts, tracks also differ greatly with respect to the opportunities they provide for students' futures, mainly through the educational credentials they offer. This feature of tracking is particularly evident when students are at the end of schooling and are getting ready to enter the labor market or further education. We thus argue that in order to gain a complete picture of the influence of tracking on students' self-beliefs, it is important to take into account the educational credentials that students receive in different tracks and analyze how they affect students' self-beliefs.

The importance of the wider social recognition attached to a person's educational credentials as a key outcome of schooling and a resource for the future is particularly emphasized in the sociology of education (Bills, 2003; Meyer, 1977). After all, employers and higher education institutions usually use previous educational credentials as criteria for selecting their employees or students. Following Bourdieu (1986), an academic qualification can be thought of as an "institutionalized objectification of cultural capital," which "confers on its holder a conventional, constant, legally guaranteed value" (p. 51). Surprisingly, the (educational) psychological research has not devoted much attention to the meaning of educational certificates for individuals. Only recently did a study by Kuppens, Easterbrook, Spears, and Manstead (2015) investigate education-based social identity and its association with well-being and social attitudes. The results clearly indicated that people identified with their level of education and that less educated people did not feel good about their level of education, which was interpreted by the authors as evidence of a social stigma. Solga (2004) has also argued that the "low education" category can be considered a social stigma in societies that are heavily dependent on human capital. In her view "less-educated youths find themselves in latent and manifest crisis situations" that are accompanied by "negative identity constructions" (p. 102). She further assumes that these students will use self-protective mechanisms, such as disengaging from school, to avoid further stigmatization. In fact, this form of disengagement can be seen as a form of social creativity in order to maintain a positive social identity, as posited by social identity theories (Kelly, 2009). Because this reasoning is also in line with the above-mentioned findings on the lack of school belonging among low track students by Van Houtte and Stevens (2009), in the present paper we explicitly investigate students' school disengagement in addition to their self-beliefs. Moreover, we believe it is important to consider self-beliefs that relate to students' future opportunities—and not only focus on academic self-concepts, which are very much tied to the school context. This is why we analyze students' self-beliefs regarding labor market entry in the present paper.

The German Tracking System

Before specifying our research questions and describing our empirical approach, it is important to provide the reader with some background information about the German tracking system[3] and to highlight why this system is an ideal context in which to study the influence of tracking on students' self-beliefs.

Germany is often used as a prototypical example of a rigid between-school tracking system. That is, students are selected into schools of different tracks at the end of elementary school, which lasts for 4 to 6 school years depending on the federal state. Even though there is considerable variation across federal states with respect to the number and quality of these school tracks, and despite the fact that some detracking reforms have taken place in recent years, Germany's traditional multitiered system of *Hauptschule*, *Realschule,* and *Gymnasium* is still evident in most states (Neumann, Becker, & Maaz, 2013).[4] The Hauptschule is the low-track school, providing a slow-paced and vocationally oriented curriculum. The Realschule is the intermediate-track school and also provides a vocational-oriented curriculum. The Gymnasium, the high-track school, provides students with an academic curriculum preparing them for higher education. In addition, there are comprehensive schools for students of all achievement levels.[5] In the present study, which uses data from the state of Berlin, we focus on low-track, intermediate-track, and comprehensive schools.

Just as there are different school tracks, there are different school-leaving certificates: the *Hauptschulabschluss* (the lowest school-leaving certificate, received either after 9th or 10th grade depending on the state), the *Mittlerer Schulabschluss* (the intermediate school-leaving certificate, received after 10th grade), and the *Abitur* (the highest school-leaving certificate, received after 12th or 13th grade). These different certificates play a crucial role in determining a person's future occupational opportunities (e.g., Protsch & Solga, 2016). The Abitur is the formal prerequisite for university enrollment. By contrast, the low and the intermediate school-leaving certificates only allow entry into the vocational educational system. The most typical form of initial vocational training, the dual apprenticeship, combines on-the-job training with education at a vocational school and can be seen as the key "entry ticket" into the labor market in Germany for nontertiary graduates (Shavit & Müller, 2000). However, even though both the low and intermediate school-leaving certificates qualify students for vocational education, it has become difficult for those with the

[3] For more information on the German school system, see Lohmar and Eckhardt (2014).

[4] Germany also has separate schools for students with special educational needs, called *Sonderschule* or *Förderschule.*

[5] Comprehensive schools, in addition to being a between-school track themselves, also practice within-school tracking. In our paper, we focus only on between-school tracks because these provide the main social context for students and determine whom students interact with on an everyday basis.

low school-leaving certificate to successfully apply for an apprenticeship, as employers increasingly prefer candidates with the intermediate school-leaving certificate or even the Abitur (Buch, Hell, & Wydra-Somaggio, 2011; C. Buchmann & Park, 2009; Kohlrausch & Solga, 2012).

In the past, school tracks and school-leaving certificates were much more intrinsically tied together: Successful students at low-track schools usually received the low-school-leaving certificate, students at intermediate-track schools received the intermediate school-leaving certificate, and students at high-track schools received the higher school-leaving certificate. For the most part, this is still true; the majority of students still finish school with the certificate traditionally connected to the curriculum level of their school track. Yet, as a result of reforms to increase the permeability of the education system, the school-leaving certificates are no longer exclusively attached to a school track, meaning that different certificates can be obtained within the same school track depending on the students' performance (see Table 1). This unique characteristic of the German tracking system makes it an ideal context to disentangle these two features of tracking—tracks as social contexts and tracks as pathways to different future opportunities—and their effects on students' self-beliefs.

The Present Study

Previous research on the BFLPE and on the influence of tracking on students' academic self-concepts, which has been a very prominent line of research within educational psychology (e.g., Liem et al., 2013; Liu et al., 2005; Marsh, 1987; Marsh & Hau, 2003; Seaton et al., 2009; Trautwein et al., 2006, 2009), has shown that being surrounded by low-achieving peers, as occurs in low tracks, makes students feel positive about their own competencies due to contrasting social comparisons. However, it has also been argued that being a member of a low track may also make a student feel negative about him- or herself as a result of assimilating social comparisons (Marsh et al., 1995, 2000; Preckel & Brüll, 2010; Trautwein et al., 2006, 2009). In most school systems around the world, contrast effects are larger than assimilation effects (see Chmielewski et al., 2013), which has led many authors to conclude that students benefit from being in a low track with respect to their academic self-concepts (Liu et al., 2005; Trautwein et al., 2006, 2009). This research on contrast and assimilation effects has taken a close look at how tracks affect students' academic self-concepts as social contexts; it has thus placed emphasis on the social mechanisms of tracking. However, tracks not only constitute distinct social contexts for students, but also provide students with different future opportunities, in particular through educational credentials. This feature of tracking, which has been particularly emphasized by sociologists, but which has not been on the educational psychological research agenda, may also have an impact on students' self-beliefs.

Based on these theoretical considerations, we aim to expand educational psychological research on the effects of tracking on students' self-beliefs by focusing on a time point in students' academic careers at which the diverging future opportunities for students become most apparent: when students are at the end of schooling and are about to enter the labor market. In addition to analyzing students' academic self-concepts in different domains, we also investigated students' self-beliefs, which relate to their future opportunities, namely students' self-beliefs regarding labor market entry. Moreover, we investigated students' school disengagement, as it has been suggested that this may be a result of low self-beliefs (Solga, 2004; Van Houtte & Stevens, 2009). Our study was conducted in the German educational system, as this context is ideally suited to studying tracking effects in general and disentangling the two features of tracking in particular. More precisely, Germany has schools of different tracks that constitute different social contexts—this is the first feature of tracking. Students also receive different school-leaving certificates, which largely determine students' future opportunities—the second feature of tracking. As these two features of tracking are no longer as intrinsically connected as they traditionally were, it is possible to disentangle their respective effects on students' self-beliefs. Due to the large differences between the German federal states regarding their tracking systems and the varying labor market conditions in different regions, both of which may bias our results, we only used data from the state of Berlin. Furthermore, we only focused on school tracks that students typically leave after 10th grade—low-track schools, intermediate-track schools, and comprehensive schools. Students graduating from these schools receive either the low or the intermediate school-leaving certificate. We excluded students attending high-track schools from our analyses, because they continue with school for 2 or 3 years longer, only then receiving their school-leaving certificate, the Abitur; this would make a comparison with students from other school tracks difficult. The dataset we used offered us a unique advantage: Instead of relying on self-reported measures of educational credentials, we could access school administrative data on the actual school-leaving certificates students received.

Table 1
School Tracks and School-Leaving Certificates in Germany

	School-leaving certificate		
School track	**Hauptschulabschluss** (low)	**Mittlerer Schulabschluss** (intermediate)	Abitur (high)
Hauptschule (low)	**x**	**x**	
Realschule (intermediate)	**x**	**x**	
Gymnasium (high)	x	x	x
Comprehensive schools	**x**	**x**	(x)

Note. (x) = Not all comprehensive schools offer the Abitur. School tracks and school-leaving certificates in bold are considered in the present study.

We sought to answer two research questions. In our first research question, we focused on the first feature of tracking and analyzed how tracks as social contexts influence students' academic self-concepts, students' self-beliefs regarding labor market entry, and students' school disengagement. In doing so, we analyzed contrast effects (operationalized via school average achievement) and assimilation effects (operationalized via track level). In line with previous studies, we expected to observe substantial contrast effects for students' academic self-concepts. With respect to the other two outcomes—students' self-beliefs regarding labor market entry and school disengagement—we did not have any specific hypotheses. As for assimilation effects, previous research indicates no or very weak assimilation effects on students' academic self-concepts for a between-school tracking system like Germany. However, it may be the case that students see their school track only as a temporary "pond" while they are at school, but use all students of their age cohort for social comparisons as soon as they leave school. This anticipated change in their reference group may make the school track they belong to more salient, resulting in larger assimilation effects (for a similar argument on college students, see Bassis, 1977). This may be even more so the case for assimilation effects on students self-beliefs regarding labor market entry. Regarding students' school engagement, we expected students in low-track schools to be more likely to disengage from school than students in intermediate-track and comprehensive schools. In our second research question, we focused on the second feature of tracking, namely students' future opportunities: We analyzed how the school-leaving certificates received by students affected their academic self-concepts, their self-beliefs regarding labor market entry, and their school disengagement over and above contrast and assimilation effects. We expected students' school-leaving certificates to be critically important for all three outcomes and to favor students with the intermediate school-leaving certificate compared to those with the low school-leaving certificate.

Method

Sample

The study draws on data from a representative sample of ninth graders in the city of Berlin, who were surveyed as part of a longitudinal study evaluating Berlin's secondary school system, the BERLIN-study (Maaz, Baumert, Neumann, Becker, & Dumont, 2013). The BERLIN-study is a joint project by the Max-Planck-Institute for Human Development (MPIB, Berlin, Germany, Principal Investigator: Prof. Dr. Jürgen Baumert), the German Institute for International Educational Research (DIPF, Frankfurt am Main/Berlin, Germany, Principal Investigator: Prof. Dr. Kai Maaz) and the Leibniz Institute for Science and Mathematics Education (IPN, Kiel, Germany, Principal Investigator: Prof. Dr. Olaf Köller). The sampling was similar to other national and international large-scale studies (e.g., PISA, see OECD, 2014): We used a two-stage random sampling procedure, first randomly sampling schools (stratified by school track) and then randomly sampling individual students within schools. The resulting representative sample consisted of 2,155 students in 87 schools including 29 low-track schools, 23 intermediate-track schools, and 35 comprehensive schools. Out of the 781 students attending a low-track school, 699 received a low school-leaving certificate, while 82 received an intermediate school-leaving certificate. Out of the 550 students attending intermediate-track schools, 147 students left school with a low school-leaving certificate and 403 with an intermediate school-leaving-certificate. As for the 824 students in comprehensive schools, 464 obtained a low school-leaving certificate and 360 an intermediate school-leaving certificate.

We used questionnaire and standardized achievement test data from two measurement points at the end of ninth grade and the end of 10th grade. The data were collected in schools by trained research assistants in May and June, 2011 and in March, 2012, respectively. Additionally, objective data on demographic variables (e.g., students' gender) and on the school-leaving certificates students obtained were collected from administrative school data after 10th grade.

Instruments

Outcome variables.

Academic self-concept. Students' academic self-concepts were assessed at the end of ninth grade on three different dimensions: mathematical, verbal, and general. The items were taken from the German version (Schwanzer, Trautwein, Lüdtke, & Sydow, 2005) of the self-description questionnaire by Marsh (1992). Each scale was comprised of four items (math: e.g., "I am good at mathematics"; verbal: e.g., "I am good at reading"; general: e.g., "Compared with others, I am not as gifted"—reversed), which students had to reply to on a 4-point-Likert scale (1 = *strongly disagree* to 4 = *strongly agree*). The internal consistencies were sufficient for all three dimensions (α_{math} = .86; α_{German} = .71; $\alpha_{general}$ = .74).

Self-beliefs regarding labor market entry. In order to investigate what self-beliefs students held regarding their chances of success on the labor market after they finished school, we focused on the dual apprenticeship market, as this is the main entry point to the labor market for young people without a university degree. The following single-item measures were specifically developed to fit our research focus: "It is difficult to get an apprenticeship position with my school-leaving certificate," "My qualifications are convincing for employers when looking for an apprenticeship position," and "I am certain that I will make a good impression during job interviews." Students were asked to rate their agreement on these three statements on a 4-point-Likert scale (1 = *strongly disagree* to 4 = *strongly agree*). The first item explicitly addressed students' beliefs regarding the value of their school-leaving certificate for finding an apprenticeship—which is the prerequisite for getting a skilled job in Germany (Shavit & Müller, 2000). The second item also looked at students' beliefs related to finding an apprenticeship, but asked about their qualifications more generally. The third item addressed students' beliefs regarding their performance at job interviews in general. Our aim was to assess specific aspects related to students' chances on the labor market instead of asking about their perceived chances more broadly. For the same reason, we did not aggregate them to a common factor, but used them as separate outcome variables. The items were administered at the end of 10th grade and thus right before students entered the apprenticeship market.

School disengagement. Students' tendency to disengage from school and scholastic activities was measured via four items (e.g.,

"If I could, I would have left school long ago") with students having to respond on a 4-point-Likert scale (1 = *strongly disagree* to 4 = *strongly agree*). The scale was based on the cynicism subscale of the well-established Maslach Burnout Inventory (Maslach, Jackson, & Leiter, 1996), which was adapted for application to the school setting. It showed good internal consistency (α = .80). The items were administered close to the end of 10th grade.

Predictor variables.

Academic achievement. Students' academic achievement was measured in ninth grade via standardized tests in three domains—mathematics, reading comprehension in German, and reading comprehension in English as a foreign language. The tests were based on the German assessments of PISA (Prenzel et al., 2007) and the national state comparison tests (Böhme et al., 2010). Mathematics was assessed with 48 items, reading comprehension in German with 28 items, and reading comprehension in English as a foreign language with 82 items. All tests were administered in a multi-matrix design, that is, each student was administered only a small subset of items (see Gonzales & Rutkowski, 2010, for more information). The tests conformed to the Rasch model and its extension as a partial-credit model, and allowed partially correct answers (Wu, Adams, Wilson, & Haldane, 2007). Weighted likelihood estimates (WLE; Warm, 1989) were used as person estimates of students' ability. IRT scaling was conducted with ConQuest 2.0 (Wu et al., 2007). Reliability was high for all three domains (mathematics: r_{eap} = .90; reading comprehension in German: r_{eap} = .89; reading comprehension in English r_{eap} = .90). School average achievement was computed from the individual student scores (see also Statistical Analyses).

School track. As mentioned above, students belonged to one of the following three school tracks: *Hauptschule*—the low-track school, *Realschule*—the intermediate track school and the comprehensive school.

Type of school-leaving certificate. In each of the three school tracks, students could obtain the same two school-leaving certificates when they left school after 10th grade: *Hauptschulabschluss*—the low school-leaving certificate, and *Mittlerer Schulabschluss*—the intermediate school-leaving certificate. Information on the certificates that students obtained came from the official report cards provided by the school officials at the end of 10th grade.

Control variables.

Gender. Information on students' gender (0 = *female*, 1 = *male*) came from administrative data reported by the school officials. 54% of students in our sample were boys.

Socioeconomic background. Students were asked to specify their parents' current occupations, which were categorized according to the International Standard Classification of Occupations (ISCO-08; ILO, 2012) and then transformed into the International Socio-Economic Index of Occupational Status (ISEI; Ganzeboom, De Graaf, & Treiman, 1992). The ISEI is a standard measure capturing a person's socioeconomic background ranging from 16 to 90, with a higher score indicating a higher status. When scores were available for both the father's and mother's occupation, the higher score was included in the analyses. In our sample, students' parents had an ISEI of M = 42.88 (SD = 18.80).

Immigrant background. A student was classified as having an immigrant background if at least one parent was born outside Germany based on the self-report data from the student questionnaire (0 = *no immigrant background*, 1 = *immigrant background*); 51.8% of students in our sample had an immigrant background.

Statistical Analyses

Handling of missing data. We gained information on demographic variables through administrative school data for all 2,155 students in our sample. However, not all students participated in the assessment. In ninth grade, 80.3% of students participated, while in 10th grade 72.5% of students participated. Overall, we had questionnaire data on at least one measurement point for 87.1% of our sample. Missing data is a notorious problem in large-scale field studies in general, and in particular for longitudinal studies. Imputation-based methods—specifically multiple imputation—are currently seen as the best way to deal with this problem: They make use of all available data, make weaker assumptions on missing data mechanisms than list- or pairwise deletion, account for the uncertainty of the value estimation, and are more robust and consistent, even when model assumptions are violated (Graham, 2009). Accordingly, we multiply imputed our data using the R package MICE ("multivariate imputation by chained equations"; van Buuren & Groothuis-Oudshoorn, 2011), thus creating 10 data sets. We integrated between- and within-imputation variance following Rubin (1987), which is automatized in Mplus through the analysis option type = imputation (Muthén & Muthén, 1998–2013).

Model estimation. We used structural equation modeling (SEM) to analyze our two research questions. Scales with multiple items (i.e., students' academic self-concepts and school disengagement) were treated as latent factors; the model fits for the respective measurement models were good: math self-concept: CFI = .99, RMSEA = .015, SRMR = .006; verbal self-concept: CFI = .99, RMSEA = .041, SRMR = .016; general academic self-concept: CFI = .99, RMSEA = .024, SRMR = .011; school disengagement: CFI = .99, RMSEA = .013, SRMR = .007. For the math and verbal self-concepts, the respective domain-specific achievement scores were used as predictors. For all other outcome variables, the mean of students' achievement scores in math, reading, and English as a foreign language was used.

For each of the outcome variables described above, three regression models were estimated, adding more predictor variables with each model. In Model 1, the outcome variables were predicted on the basis of individual achievement, school mean achievement, and track level. Since there were three school tracks, we used two dummy variables (intermediate-track school and comprehensive school), with the low-track school being the reference category. This model allowed us to analyze contrast and assimilation effects, which were the focus of our first research question. In Model 2, we analyzed our second research question by adding students' obtained school-leaving certificates to the model. For better interpretability, instead of entering them into the model as interaction terms, we created dummy variables indicating both the school-leaving certificate students obtained and the school track students belonged to. This resulted in five dummy variables: low school-leaving certificate at an intermediate-track school, low school-leaving certificate at a comprehensive school,

intermediate school-leaving certificate at a low-track school, intermediate school-leaving certificate at an intermediate-track school, and intermediate school-leaving certificate at a comprehensive school, with students expecting a low school-leaving certificate and attending a low-track school as the reference group. To ensure the robustness of the results, we also included gender, SES, and immigrant background as control variables in Model 3.

All models were estimated in Mplus 7.1 (Muthén & Muthén, 1998–2013) using the complex sampling option (type = complex), in which standard errors are corrected to account for the multilevel structure of the data (using schools as the clustering variable). We decided to use this approach instead of multilevel modeling because it permits the use of dummy indicators for combinations of school track and school-leaving certificates (as described above) to facilitate interpretation. Sampling weights on both the school and the individual level were used to account for differential sampling probabilities in order to ensure the representativeness of our sample.

Results

The correlations of all variables considered in the analyses are presented in Table 2. The results of the three models addressing our research questions can be found in Table 3. In the following, we will describe the findings for each outcome separately, that is, for students' academic self-concepts, students' self-beliefs regarding labor market entry, and students' school disengagement. For all latent models, the model fit indices were below the established cut-off criteria (CFI > .95; RMSEA < .06; SRMR < .08; Hu & Bentler, 1999).

Students' Academic Self-Concept

We first looked at the prediction of students' academic self-concepts on the basis of individual achievement, school mean achievement, and track level in order to analyze contrast and assimilation effects (Model 1), and found relatively consistent results for all three domains: Students' academic self-concepts were not only shaped by their individual achievement but also by the achievement of their schoolmates—as indicated by the statistically negative coefficients of school mean achievement, showing that the higher the average achievement level at a school, the lower students' academic self-concepts net of their individual achievement (math: $\beta = -.11$, $p < .05$; verbal: $\beta = -.16$, $p < .01$; general: $\beta = -.12$, $p < .05$). Therefore, as predicted, contrast effects were present for all three domains. Regarding the influence of the track level, we found no statistically significant differences in students' math and general academic self-concepts, indicating that we could not observe any positive assimilation effects of belonging to a higher track. Only for the verbal domain did we find that students attending intermediate-track schools had a statistically significantly higher academic self-concept than students attending low-track schools ($\beta = .31$, $p < .05$). No differences were found between low-track and comprehensive schools. Taken together, the contrast effects of school mean achievement played a much larger role in predicting students' academic self-concepts than the assimilation effects of the track level, which is in line with our hypotheses.

Regarding the influence of the school-leaving certificates that students obtained at the end of 10th grade (Model 2), students with an intermediate school-leaving certificate had higher academic self-concepts in all three domains than students expecting a low school-leaving certificate, no matter which school track they belonged to and independent of their own and their school's mean achievement (math: Intermediate SLC at low-track school: $\beta = .35$, $p < .01$, Intermediate SLC at intermediate-track school: $\beta = .29$, $p < .05$, Intermediate SLC at comprehensive school: $\beta = .32$, $p < .01$; verbal: Intermediate SLC at low-track school: $\beta = .63$, $p < .001$, Intermediate SLC at intermediate-track school: $\beta = .54$, $p < .001$, Intermediate SLC at

Table 2
Correlations of All Variables Considered in the Present Study

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Individual achievement																
2. School mean achievement	*.65*															
3. Low SLC at intermediate-track school	*−.05*	*.10*														
4. Low SLC at comprehensive school	*−.12*	*.05*	*−.16*													
5. Intermediate SLC at low-track school	−.03	*−.20*	*−.05*	*−.09*												
6. Intermediate SLC at intermediate-track school	*.29*	*.30*	*−.17*	*−.32*	*−.09*											
7. Intermediate SLC at comprehensive school	*.36*	*.32*	*−.15*	*−.29*	*−.08*	*−.30*										
8. Sex (1 = boy)	*−.08*	*−.07*	−.02	.00	.00	−.02	−.04									
9. Socioeconomc background	*.28*	*.31*	*−.08*	−.03	−.01	.05	*.21*	.02								
10. Immigrant background	*−.24*	*−.26*	.06	*.06*	−.02	−.03	*−.10*	.00	*−.17*							
11. Math self-concept	*.22*	*.11*	−.05	*−.15*	*.05*	*.12*	*.15*	*.24*	.07	−.03						
12. Verbal self-concept	*.23*	*.07*	−.05	*−.10*	*.08*	*.12*	*.12*	*−.15*	*.14*	−.01	−.01					
13. General academic self-concept	*.31*	*.15*	−.06	*−.14*	.05	*.17*	*.14*	*.19*	*.17*	−.05	*.31*	*.36*				
14. Self-belief regarding labor market entry: It is difficult . . .	*−.31*	*−.23*	.07	*.21*	−.04	*−.17*	*−.25*	.06	*−.16*	*.07*	*−.24*	*−.22*	*−.25*			
15. Self-belief regarding labor market entry: My qualifications . . .	*.20*	*.13*	−.05	*−.11*	.03	*.07*	*.16*	*−.06*	*.17*	−.04	*.11*	*.22*	*.20*	*−.21*		
16. Self-belief regarding labor market entry: I am certain . . .	*.16*	*.13*	−.02	−.05	.04	*.08*	*.07*	−.05	*.16*	−.06	.04	*.23*	*.15*	*−.14*	*.48*	
17. School disengagement	*−.20*	*−.12*	*.08*	*.17*	*−.06*	*−.11*	*−.21*	*.08*	*−.12*	−.02	*−.19*	*−.26*	*−.23*	*.32*	*−.17*	*−.12*

Note. SLC = School-leaving certificate; Statistically significant correlations are shown in italics; correlations involving scales with multiple items (i.e., students' academic self-concepts and school disengagement) and individual and school mean achievement all represent latent correlations.

Table 3
Predicting Students' Academic Self-Concepts, Self-Beliefs Regarding Labor Market Entry, and School Disengagement

Models	Academic self-concept						Self-beliefs regarding labor market entry							
	Math self-concept		Verbal self-concept		General academic self-concept		"It is difficult to receive an apprenticeship with my school degree."		"My qualifications are convincing when looking for apprenticeship."		"I am certain that I will make a good impression during job interviews."		School disengagement	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Model 1														
Individual achievement	.38***	.03	.26***	.04	.36***	.04	−.28***	.04	.20***	.04	.14**	.05	−.21***	.04
School mean achievement	−.11*	.04	−.16**	.06	−.12*	.06	−.06	.05	.02	.05	.06	.05	.02	.05
Intermediate-track school	.10	.06	.31*	.12	.19	.12	−.01	.11	−.10	.11	−.03	.12	−.02	.11
Comprehensive school	.01	.10	.22	.14	.04	.10	.09	.10	−.09	.10	−.09	.10	.02	.09
R^2	.12		.06		.11		.10		.04		.03		.04	
Model 2														
Individual achievement	.34***	.04	.21***	.04	.30***	.04	−.17***	.04	.15**	.05	.11*	.05	−.10*	.05
School mean achievement	−.14**	.04	−.19**	.06	−.13*	.06	−.02	.05	.01	.05	.06	.05	.05	.06
Low SLC at intermediate-track school	.05	.14	.29	.16	.10	.16	.05	.14	−.11	.13	−.05	.14	.04	.16
Low SLC at comprehensive school	−.04	.10	.21	.14	.03	.11	.18	.09	−.14	.10	−.07	.11	.08	.10
Intermediate SLC at low-track school	.35**	.13	.63***	.16	.40*	.217	−.44**	.16	.23	.16	.33*	.13	−.54**	.20
Intermediate SLC at intermediate-track school	.29*	.12	.54***	.15	.40**	.13	−.37**	.13	.08	.12	.07	.12	−.37**	.14
Intermediate SLC at comprehensive school	.32**	.12	.57**	.17	.33**	.13	−.52***	.13	.23	.13	.03	.11	−.55***	.15
R^2	.14		.09		.12		.16		.06		.03		.09	
Model 3														
Individual achievement	.31***	.04	.19***	.05	.30***	.04	−.16***	.04	.13**	.05	.10	.05	−.10*	.05
School mean achievement	−.11*	.05	−.21**	.06	−.15**	.06	−.01	.06	−.02	.06	.02	.05	.03	.06
Low SLC at intermediate-track school	.06	.13	.30	.16	.17	.16	.04	.14	−.09	.14	.00	.15	.11	.16
Low SLC at comprehensive school	−.04	.10	.20	.14	.06	.11	.19	.10	−.14	.10	−.05	.11	.14	.10
Intermediate SLC at low-track school	.40**	.14	.59***	.16	.41*	.19	−.42*	.16	.20	.16	.31*	.13	−.53**	.20
Intermediate SLC at intermediate-track school	.31**	.12	.54***	.15	.43**	.13	−.36**	.13	.08	.12	.09	.12	−.30*	.13
Intermediate SLC at comprehensive school	.36**	.12	.53**	.18	.35**	.13	−.50***	.13	.20	.12	.02	.11	−.47***	.14
Sex (1 = boy)	.45***	.05	−.25***	.06	.41***	.06	.08	.07	−.10*	.05	−.09	.06	.13	.07
Socioeconomc background	−.01	.04	.12**	.04	.10**	.04	−.06	.05	.12**	.05	.12***	.03	−.07	.04
Immigrant background	.06	.07	.03	.08	.03	.06	−.03	.07	.04	.07	−.01	.07	−.16*	.07
R^2	.19		.11		.17		.17		.07		.05		.11	

Note. SLC = School-leaving certificate; all continuous variables were standardized beforehand.
* $p < .05$. ** $p < .01$. *** $p < .001$.

comprehensive school: β = .57, $p < .01$; general: Intermediate SLC at low-track school: β = .40, $p < .05$, Intermediate SLC at intermediate-track school: β = .40, $p < .01$, Intermediate SLC at comprehensive school: β = .33, $p < .01$). These results were robust even when we added students' gender, parental SES, and immigrant status as further predictors (Model 3). Therefore, our hypothesis that students anticipating the low school-leaving certificate have a lower academic self-concept than students expecting to obtain an intermediate school-leaving certificate was confirmed.

Students' Self-Beliefs Regarding Labor-Market Entry

Turning to students' self-beliefs regarding labor market entry, which were measured via three single items, we did not find any evidence for contrast or assimilation effects. That is, school mean achievement and the school track students belonged to did not serve as statistically significant predictors after controlling for individual achievement, which did serve as a predictor for all three items (see results for Model 1). As for the influence of school-leaving certificates (Model 2), students with intermediate school-leaving certificates believed that they would find it less difficult to secure an apprenticeship position with their school-leaving certificate compared to students with low school-leaving certificates—independent of their individual achievement, their school's mean achievement, and the school track they belonged to (Intermediate SLC at low-track school: β = −.44, $p < .01$; Intermediate SLC at intermediate-track school: β = −.37, $p < .01$, Intermediate SLC at comprehensive school: β = −.52, $p < .001$). With respect to the item "My qualifications are convincing for employers when looking for an apprenticeship," we did not find any differences between students with different school-leaving certificates. As for students' belief regarding the impression they would make during a job interview, students who obtained an intermediate school-leaving certificate while attending a low-track school had a statistically significantly higher self-belief than students going to a low-track school who obtained a low school-leaving certificate (β = .33, $p < .05$). No other statistically significant differences were found. Our hypothesis on students' self-beliefs regarding their labor market entry was thus only partially confirmed. The findings stayed the same when the control variables were entered into the model (Model 3). Interestingly, students' socioeconomic background turned out to be a statistically significant predictor for the last two items, indicating that students from more privileged backgrounds had higher self-beliefs (controlling for all other predictors).

Students' School Disengagement

In this case, too, we did not find any evidence for contrast or assimilation effects on students' school disengagement. This was indicated by the fact that school mean achievement and track level did not serve as statistically significant predictors after controlling for individual achievement, which negatively predicted students' school disengagement (Model 1). Whereas we did not expect to find contrast effects, contrary to our hypothesis, the school track students belonged to did not matter for their school disengagement. However, we did find large differences between students with different school-leaving certificates: Model 2 shows that students with an intermediate school-leaving certificate were much less likely to disengage from school than students with a low school-leaving certificate (Intermediate SLC at low-track school: β = −.54, $p < .01$; Intermediate SLC at intermediate-track school: β = −.37, $p < .01$; Intermediate SLC at comprehensive school: β = −.55, $p < .001$). These findings remained robust when the control variables were entered (Model 3). Therefore, our hypothesis that students with a low certificate disengage from school was confirmed.

In order to illustrate the main findings for all outcome variables considered, we depict the regression coefficients from Model 3 for the six different groups of students in Figure 1 (students' academic

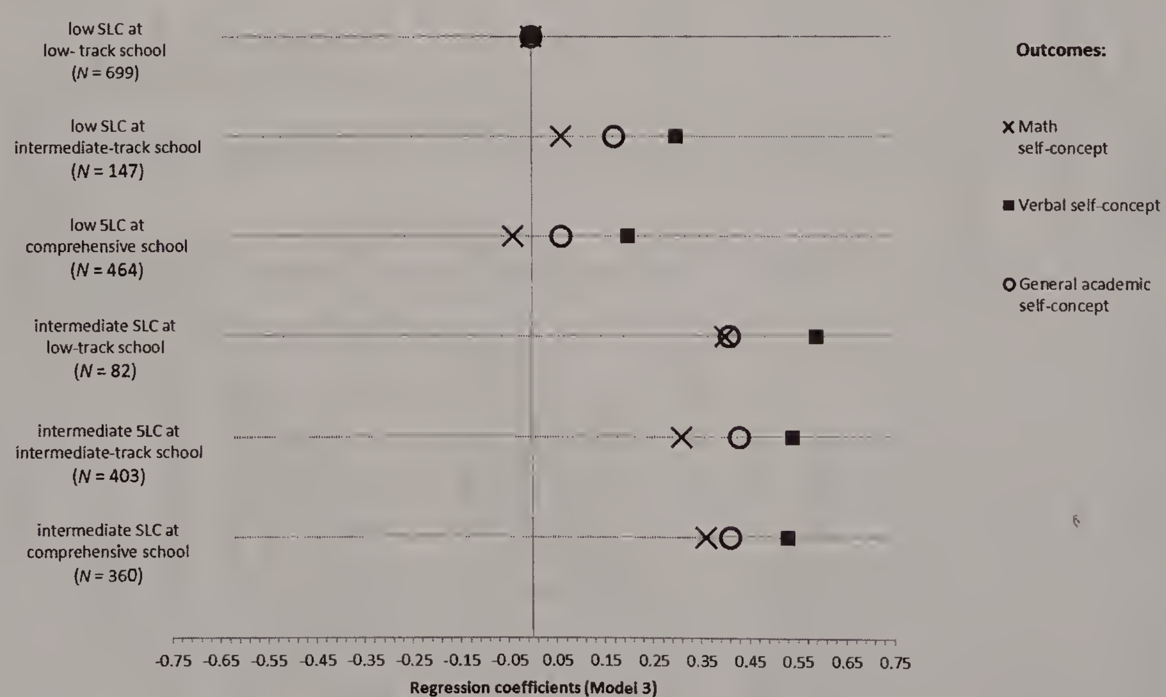


Figure 1. Regression coefficients for the school track/school-leaving-certificate categories from Model 3 for students' academic self-concepts *Note.* SLC = School-leaving certificate; the category "low SLC at low-track school" represents the reference category.

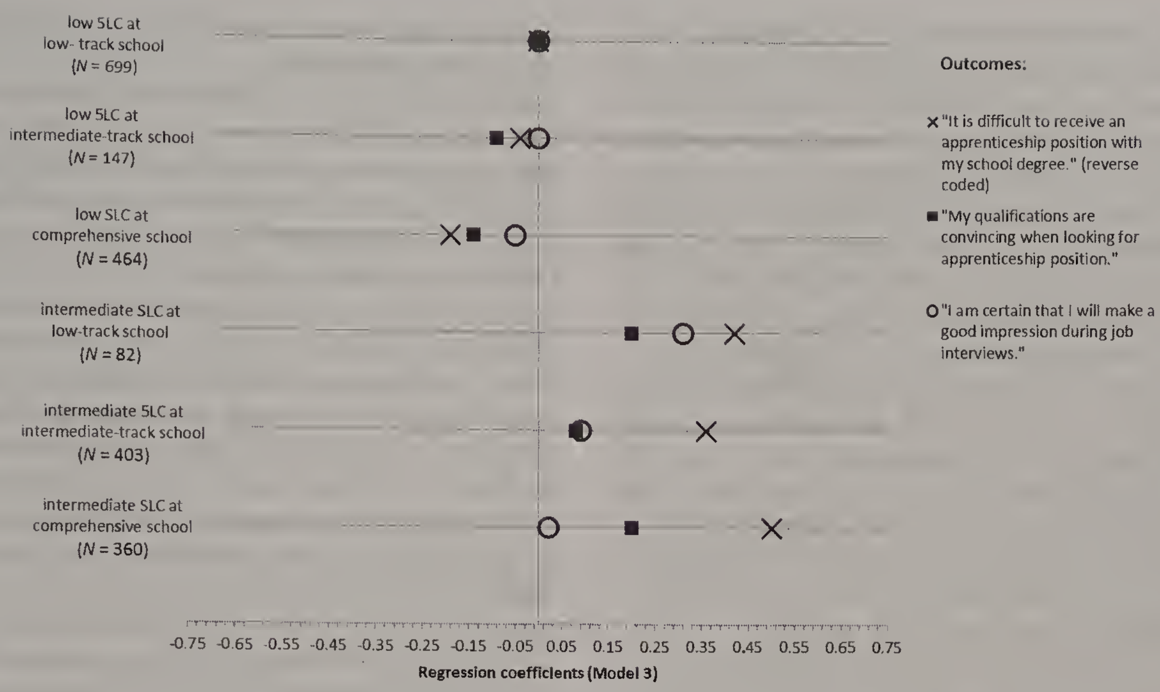


Figure 2. Regression coefficients for the school track/school-leaving-certificates categories from Model 3 for students' self-beliefs regarding labor market entry. SLC = School-leaving certificate; the category "low SLC at low-track school" represents the reference category. To facilitate interpretation, reversed-coded items were coded in the same direction.

self-concepts), Figure 2 (students' self-beliefs regarding labor market entry), and Figure 3 (students' school disengagement). These clearly show that systematic differences in student outcomes exist between students with different school-leaving certificates and not between students attending schools of different tracks.

Discussion

In the present paper, we analyzed how tracking relates to students' self-beliefs (students' academic self-concepts in different domains and their self-beliefs regarding labor-market entry) and students' school disengagement during a time period that has received little attention in the educational psychological research on tracking: when students are at the end of schooling and on the verge of entering the labor market. In doing so, we aimed to disentangle the effects of two distinguishing features of tracking: On the one hand, tracks constitute distinct social contexts for students, and on the other hand, tracks provide students with different future opportunities. Whereas the first feature has been

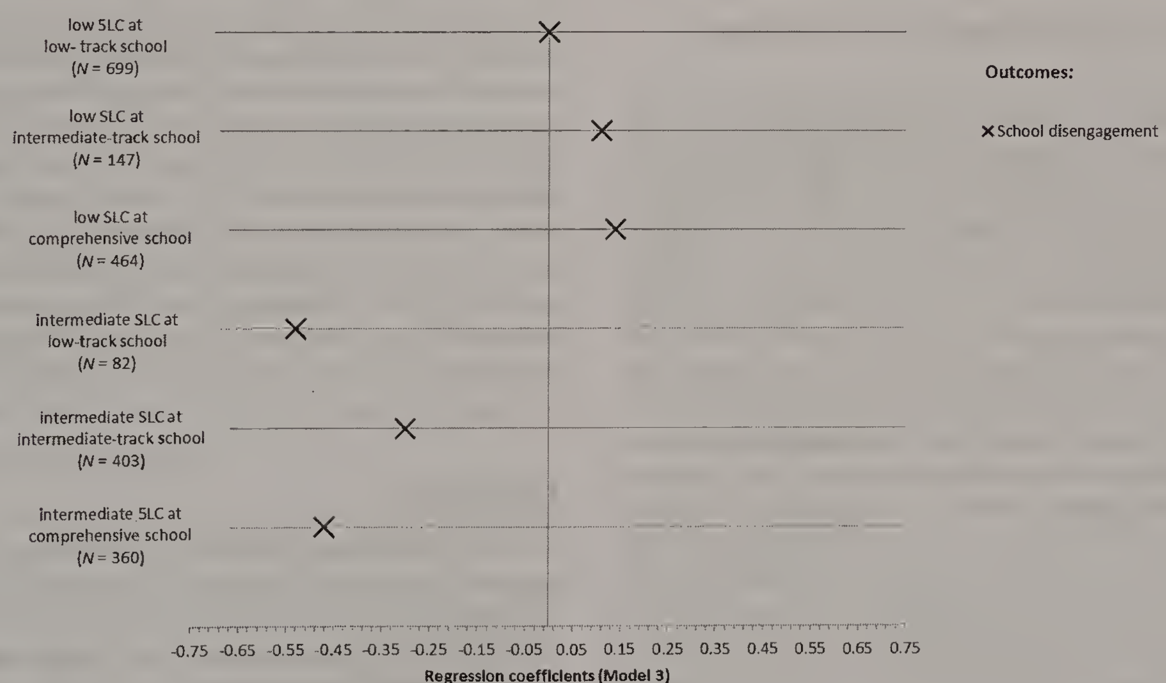


Figure 3. Regression coefficients for the school track/school-leaving-certificates categories from Model 3 for students' school disengagement. SLC = School-leaving certificate; the category "low SLC at low-track school" represents the reference category.

studied within the BFLPE research (e.g., Chmielewski et al., 2013; Preckel & Brüll, 2010; Trautwein et al., 2006), the second feature, which is much more prominent in sociology, has not yet received much attention in studies on the effects of tracking on students' self-beliefs. Our aim was thus to extend previous research in educational psychology by bringing the sociological perspective into play. We used data from the state of Berlin, Germany, which provided an ideal context to disentangle these two features of tracking. This is because Germany has schools of different tracks that constitute different social contexts—the first feature of tracking. Students also receive different school-leaving certificates, which largely determine students' future opportunities—the second feature of tracking. But even though school-leaving certificates are associated with the curriculum of a particular school track, it is possible to receive the low and intermediate school-leaving certificates at different school tracks: at low-track schools, intermediate-track schools, and comprehensive schools. This allowed us to disentangle the two features of tracking.

Overall, our results clearly point to the importance of educational certificates and thus the second feature of tracking in shaping students' self-beliefs and school disengagement. Irrespective of their individual achievement, their schoolmates' achievement, and their track level, students who received the intermediate school-leaving certificate had higher academic self-concepts, believed they would have better chances of success on the labor market with their certificate, and were less disengaged from school than students who received the low school-leaving certificate. In contrast, the school track students belonged to did not serve as a predictor for the outcomes considered. Thus, no assimilation effects could be observed. Contrast effects could only be shown for students' academic self-concepts (thus replicating the BFLPE), but not for students' self-beliefs regarding labor market entry or their school disengagement. In the following, we will discuss our findings in more depth, address the limitations of our study, and outline potential avenues for future research.

Our Findings in Light of Research on Tracking Effects

Regarding the effects of tracking on students' self-beliefs, there is a large consensus in educational psychology that being surrounded by lower achieving students, as is the case in low tracks, has positive consequences for students' academic self-concepts because students have fewer chances for upward social comparisons (Liu et al., 2005; Marsh et al., 1995; Trautwein et al., 2006, 2009). Our study shows that this may be only part of the story. First, there are a number of other self-beliefs in addition to academic self-concept; these have received far less attention but may also be relevant to look at. In particular, when considering the time point when students are about to leave school, other types of self-beliefs may become more important than subject-specific competence beliefs. In our study, we did not find any positive consequences of being surrounded by low achieving peers for students' self-beliefs regarding labor-market entry. There is reason to think that this may also be true for other self-beliefs. For instance, research looking at student's career-related self-concepts (Fuligni et al., 1995) and students' global self-esteem (Van Houtte et al., 2012) has found negative effects of belonging to a lower school track. Therefore, a sole focus on students' academic self-concepts may be too limited to fully assess the effects of tracking on students' self-beliefs.

Second, tracks are not only characterized by the social contexts they provide students. Our study shows that it is important to also take into account how tracks influence students' future trajectories; this influence often becomes visible through different educational credentials or certificates. Based on our findings, there is reason to believe that students identify with the level of education they have received, which leads students with a low educational certificate to feel less competent and to disengage from school. Even though both school-leaving certificates could be obtained in all three school tracks in our sample, most students left school with the certificate associated with the school track they belonged to. Therefore, researchers aiming to study tracking effects in a particular educational context should take into account how strongly a track determines the educational certificate or credentials a student will receive at the end of schooling. This is particularly relevant for school systems like Germany, in which different school-leaving certificates exist. However, similar features can be found in all tracking systems. For instance, in the United States, highly selective colleges consider courses taken in high school, including the number of college-preparatory and AP classes, and the AP test scores, when admitting students. Thus, in a more informal way, high school transcripts and test scores in the United States may constitute educational credentials.

Implications for Research on the BFLPE

The findings of our study also have implications for research on the BFLPE. With more than 70 articles published in leading APA journals in just the past decade, the BFLPE is one of the most prominent phenomena in educational psychology and has even been called a "pan-human theory" (Seaton et al., 2009). In fact, the BFLPE has been shown to persist after high school (Marsh et al., 2007) and to generalize to outcomes other than academic self-concept (Marsh, 1991; Marsh & O'Mara, 2010; Nagengast & Marsh, 2012). We think our results strengthen Dai's (2004) argument that "the BFLPE is only part of this much larger story about personal and academic development" (p. 303) by showing that the educational certificates students obtained were at least as important for their academic self-concepts as the academic achievement of their peers. Also, we did not find a negative coefficient for either school mean achievement on students' self-beliefs regarding labor market entry or for students' school disengagement. This implies that even though students in lower tracks, who are surrounded by low achieving peers, may feel positive about their competence in school subjects such as math and reading, they still hold realistic beliefs about their lower chances on the labor market.

Limitations

There are several limitations of our study that need to be addressed. In terms of students' self-beliefs beyond academic self-concept, we only focused on students' self-beliefs regarding their immediate chances on the labor market when looking for an apprenticeship. There may well be other self-beliefs worth looking at in order to fully understand how students perceive their own academic standing and their future chances at the end of schooling. Moreover, we had to rely on single item measures for students'

self-beliefs regarding labor market entry. This makes our results less reliable and thus represents a clear limitation of our study. Hence, it seems worthwhile for future studies to develop multi-item measures to more accurately assess students' future-directed self-beliefs. Additionally, the variables analyzed in the present study were measured at different time points. Whereas students' academic self-concepts were measured at the end of ninth grade, students' self-beliefs regarding labor market entry and their disengagement from school were measured in 10th grade. We do not know whether and how these different measurement points influenced our findings, which is a clear limitation of our study. For instance, it is possible that students feel more disengaged from school in 10th grade than in ninth grade because they are tired of school. Last but not least, our data set does not allow for causal inference and we can only make theoretical assumptions about the mechanisms underlying our findings. We interpret our findings to mean that students anticipate receiving a certain certificate, which then has an influence on their self-beliefs and the way they engage in school. However, one could also argue that a lower academic self-concept and a high level of school disengagement might lead to different educational attainment. While we cannot rule out these effects, we did control for students' achievement and do not believe that differences in self-beliefs and school disengagement can fully explain the school-leaving certificates obtained. Most likely, the association we observed between students' obtained school-leaving certificates and their self-beliefs/school disengagement may represent a "vicious circle": For instance, anticipating a low school-leaving certificate may result in a lower self-belief and disengagement from school, which then in turn makes it even more likely that a student will receive this certificate. Similarly, a student expecting to finish school with the intermediate school-leaving certificate may have higher self-beliefs and may be more engaged in school, which then increases his or her chances of actually receiving this certificate. Like all other tracking research, our study faces the problem of causality on a more general level—track assignment is nonrandom by definition and often produces groups with differing socioeconomic compositions. Therefore, a very cautious interpretation of our findings is that they could also be a result of selection into tracks. That is to say, it could be the case that students in the low track were already low on the considered outcome measures. However, we not only considered how students' self-beliefs and school-disengagement differed according to the track they belonged to; we also looked at variation within tracks by analyzing a school's mean achievement and students' school-leaving certificates. More generally, the fact that our findings were robust after adding control variables to our model also speaks to their validity.

Avenues for Future Research

Finally, we would like to outline several avenues for future research in response to our study. First, we would like to see the educational psychological research agenda on tracking moving beyond the BFLPE. As has been argued by Van Houtte and Stevens (2009), who stated that "future research should deal explicitly with relative deprivation and stigmatizing effects of being tracked and the stigmatizing character of certain school types" (p. 964), this should include an in depth-analysis of stigmatization and labeling effects. Our study can be seen as a first step in this direction. By disentangling the effect of school track and educational certificates, we were able to show that it is not the track level and therefore the label of the track, but rather the educational certificates associated with a school track that lead to stigmatizing effects. In a similar vein, we believe that Karlson's (2015) finding that "adolescents actively revise their educational expectations in response to their track placement in high school—an ability signal whose value . . . derives from its relation to adolescents' perceived chances for success in future schooling" may actually be driven by the educational credentials students typically receive in different tracks. We encourage tracking researchers to pay attention to the fact that tracks may not only constitute social contexts, but also contribute to inequalities in future educational and occupational career paths and related self-beliefs by awarding different educational credentials. Second, it would be worthwhile to examine how expectations on the part of teachers and others may influence students' academic self-beliefs and their engagement in school. For instance, previous research has shown that teachers expect less from students in lower tracks (e.g., Kelly & Carbonaro, 2012), which could negatively affect students' own perceptions of their competencies and future chances. Third, in line with Kuppens et al. (2015), our findings also suggest that an individual's level of education is an important part of his or her social identity. Analyzing the meaning of people's education-based social identities may be a very promising research avenue within educational psychology. Fourth, on a more general note, our study is an example of how helpful the integration of other disciplinary perspectives can be. We believe that by bringing the sociological perspective into play, we were able to offer a more complete picture of the consequences of tracking for students' self-beliefs. Finally, longitudinal studies that go beyond students' time at school and follow them into their working lives are needed to clarify the long-term role of students' self-beliefs for their occupational trajectories.

References

Alicke, M. D., Zell, E., & Bloom, D. L. (2010). Mere categorization and the frog-pond effect. *Psychological Science, 21,* 174–177. http://dx.doi.org/10.1177/0956797609357718

Allmendinger, J. (1989). Educational systems and labor market outcomes. *European Sociological Review, 5,* 231–250.

Bassis, M. S. (1977). The campus as a frog pond: A theoretical and empirical reassessment. *American Journal of Sociology, 82,* 1318–1326. http://dx.doi.org/10.1086/226467

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology, 104,* 682–699. http://dx.doi.org/10.1037/a0027608

Bills, D. B. (2003). Credential, signals, and screens: Explaining the relationship between schooling and job assignment. *Review of Educational Research, 73,* 441–449. http://dx.doi.org/10.3102/00346543073004441

Böhme, K., Leucht, M., Schipolowski, S., Porsch, R., Knigge, M., & Köller, O. (2010). Anlage und Durchführung des Ländervergleichs. [Design of the National Assessment Study]. In O. Köller, M. Knigge, & B. Tesch (Eds.), *Sprachliche Kompetenzen im Ländervergleich* [Language skills in the National Assessment Study] (pp. 65–85). Münster, Germany: Waxmann.

Bol, T., & Van de Werfhorst, H. G. (2013). Educational system and the trade-of between labor market allocation and equality of educational

opportunity. *Comparative Education Review, 57,* 285–308. http://dx.doi.org/10.1086/669122

Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook of theory and research for the sociology of education* (pp. 241–260). New York, NY: Greenwood Press.

Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy, 22,* 782–861. http://dx.doi.org/10.1111/j.1468-0327.2007.00189.x

Brzinsky-Fay, C. (2007). Lost in transition? Labour market entry sequences of school leavers in Europe. *European Sociological Review, 23,* 409–422. http://dx.doi.org/10.1093/esr/jcm011

Buch, T., Hell, S., & Wydra-Somaggio, G. (2011). Stigma hauptschulabschluss? Der einfluss der schulbildung auf das arbeitslosigkeitsrisiko an der zweiten schwelle [The stigma of lower school qualifications. The influence of school education on the probability of unemployment]. *Zeitschrift für Erziehungswissenschaft, 14,* 421–443. http://dx.doi.org/10.1007/s11618-011-0214-3

Buchmann, C., & Park, H. (2009). Stratification and the formation of expectations in highly differentiated educational systems. *Research in Social Stratification and Mobility, 27,* 245–267. http://dx.doi.org/10.1016/j.rssm.2009.10.003

Buchmann, M., & Kriesi, I. (2011). Transition to adulthood in Europe. *Annual Review of Sociology, 37,* 481–503. http://dx.doi.org/10.1146/annurev-soc-081309-150212

Carbonaro, W. (2005). Tracking, students' effort, and academic achievement. *Sociology of Education, 78,* 27–49. http://dx.doi.org/10.1177/003804070507800102

Chiu, D., Beru, Y., Watley, E., Wubu, S., Simson, E., Kessinger, R., . . . Wigfield, A. (2008). Influences of math tracking on seventh-grade students' self-beliefs and social comparisons. *The Journal of Educational Research, 102,* 125–136. http://dx.doi.org/10.3200/JOER.102.2.125-136

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type: An international comparison of students' mathematics self-concept. *American Educational Research Journal, 50,* 925–957. http://dx.doi.org/10.3102/0002831213489843

Cialdini, R. B., Borden, R. J., Thorne, A., Walker, M. R., Freeman, S., & Sloan, L. R. (1976). Basking in reflected glory: Three (football) field studies. *Journal of Personality and Social Psychology, 34,* 366–375. http://dx.doi.org/10.1037/0022-3514.34.3.366

Dai, D. Y. (2004). How universal is the big-fish-little-pond effect? *American Psychologist, 59,* 267–268. http://dx.doi.org/10.1037/0003-066X.59.4.267

Fuligni, A. J., Eccles, J. S., & Barber, B. L. (1995). The long-term effects of seventh-grade ability grouping in mathematics. *The Journal of Early Adolescence, 15,* 58–89. http://dx.doi.org/10.1177/0272431695015001005

Gamoran, A. (1986). Instructional and institutional effects of ability grouping. *Sociology of Education, 59,* 185–198. http://dx.doi.org/10.2307/2112346

Gamoran, A. (1992). Is ability grouping equitable? *Educational Leadership, 50,* 11–17.

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21,* 1–56. http://dx.doi.org/10.1016/0049-089X(92)90017-B

Gonzales, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 3,* 125–156.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530

Heckhausen, J., Chang, E. S., Greenberger, E., & Chen, C. (2013). Striving for educational and career goals during the transition after high school: What is beneficial? *Journal of Youth and Adolescence, 42,* 1385–1398. http://dx.doi.org/10.1007/s10964-012-9812-5

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Huguet, P., Dumas, F., Marsh, H., Régner, I., Wheeler, L., Suls, J., . . . Nezlek, J. (2009). Clarifying the role of social comparison in the big-fish-little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology, 97,* 156–170. http://dx.doi.org/10.1037/a0015558

ILO. (2012). *International standard classification of occupations* (ISCO-08). Geneva, Switzerland: Author.

Karlson, K. B. (2015). Expectations on track? High school tracking and adolescent educational expectations. *Social Forces, 94,* 115–141. http://dx.doi.org/10.1093/sf/sov006

Kelly, S. (2009). Social identity theories and educational engagement. *British Journal of Sociology of Education, 30,* 449–462. http://dx.doi.org/10.1080/01425690902954620

Kelly, S., & Carbonaro, W. (2012). Curriculum tracking and teacher expectations: Evidence from discrepant course taking models. *Social Psychology of Education, 15,* 271–294. http://dx.doi.org/10.1007/s11218-012-9182-6

Knigge, M., & Hannover, B. (2011). Collective school-type identity: Predicting students' motivation beyond academic self-concept. *International Journal of Psychology, 46,* 191–205. http://dx.doi.org/10.1080/00207594.2010.529907

Kohlrausch, B., & Solga, H. (2012). Übergänge in die ausbildung: Welche rolle spielt die ausbildungsreife [Transitions into apprenticeship in Germany: How important is youth's "maturity for VET"?]. *Zeitschrift für Erziehungswissenschaft, 15,* 753–773. http://dx.doi.org/10.1007/s11618-012-0332-6

Kuppens, T., Easterbrook, M. J., Spears, R., & Manstead, A. S. R. (2015). Life at both ends of the ladder: Education-based identification and its association with well-being and social attitudes. *Personality and Social Psychology Bulletin, 41,* 1260–1275. http://dx.doi.org/10.1177/0146167215594122

Liem, G. A. D., Marsh, H. W., Martin, A. J., McInerney, D. M., & Yeung, A. S. (2013). The big-fish-little-pond effect and a national policy of within-school ability streaming: Alternative frames of reference. *American Educational Research Journal, 50,* 326–370. http://dx.doi.org/10.3102/0002831212464511

Liu, W. C., Wang, C. K. J., & Parkins, E. J. (2005). A longitudinal study of students' academic self-concept in a streamed setting: The Singapore context. *British Journal of Educational Psychology, 75,* 567–586. http://dx.doi.org/10.1348/000709905X42239

Lohmar, B., & Eckhardt, T. (2014). *The education system in the Federal Republic of Germany 2012/2013.* Bonn, Germany: Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs.

Lucas, S. R. (1999). *Tracking inequality: Stratification and mobility in American high schools.* New York, NY: Teachers College Press.

Lucas, S. R., & Berends, M. (2002). Sociodemographic diversity, correlated achievement, and de facto tracking. *Sociology of Education, 75,* 328–348. http://dx.doi.org/10.2307/3090282

Maaz, K., Baumert, J., Neumann, M., Becker, M., & Dumont, H. (2013). *Die Berliner Schulstrukturreform. Bewertung durch die beteiligten Akteure und Konsequenzen des neuen Übergangsverfahrens von der Grundschule in die weiterführenden Schulen* [The secondary school reform in Berlin]. Münster, Germany: Waxmann.

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational

outcomes. *Child Development Perspectives, 2,* 99–106. http://dx.doi.org/10.1111/j.1750-8606.2008.00048.x

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79,* 280–295. http://dx.doi.org/10.1037/0022-0663.79.3.280

Marsh, H. W. (1990). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review, 2,* 77–172. http://dx.doi.org/10.1007/BF01322177

Marsh, H. W. (1991). Failure of high-ability high schools to deliver academic benefits commensurate with their students' ability levels. *American Educational Research Journal, 28,* 445–480. http://dx.doi.org/10.3102/00028312028002445

Marsh, H. W. (1992). *Self Description Questionnaire (SDQ) III: A theoretical and empirical basis for the measurement of multiple dimensions of late adolescent self-concept* (An interim test manual and a research monograph). Macarthur, New South Wales, Australia: University of Western Sydney, Faculty of Education.

Marsh, H. W., Chessor, D., Craven, R., & Roche, L. (1995). The effect of gifted and talented programs on academic self-concept: The big fish strikes again. *American Educational Research Journal, 32,* 285–319. http://dx.doi.org/10.3102/00028312032002285

Marsh, H. W., & Hau, K. T. (2003). Big-fish-little-pond effect on academic self-concept. A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist, 58,* 364–376. http://dx.doi.org/10.1037/0003-066X.58.5.364

Marsh, H. W., Hau, K-T., & Craven, R. G. (2004). The big-fish-little-pond effect stands up to scrutiny. *American Psychologist, 59,* 269–271. http://dx.doi.org/10.1037/0003-066X.59.4.269

Marsh, H. W., Kong, C. K., & Hau, K. T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology, 78,* 337–349. http://dx.doi.org/10.1037/0022-3514.78.2.337

Marsh, H. W., & O'Mara, A. (2008). Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: Unidimensional and multidimensional perspectives of self-concept. *Personality and Social Psychology Bulletin, 34,* 542–552. http://dx.doi.org/10.1177/0146167207312313

Marsh, H. W., & O'Mara, A. J. (2010). Long-term total negative effects of school-average ability on diverse educational outcomes. *Zeitschrift für Pädagogische Psychologie, 24,* 51–72. http://dx.doi.org/10.1024/1010-0652/a000004

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K-T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish-little-pond effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review, 20,* 319–350. http://dx.doi.org/10.1007/s10648-008-9075-6

Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). The big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal, 44,* 631–669. http://dx.doi.org/10.3102/0002831207306728

Maslach, C., Jackson, S. E., & Leiter, M. P. (1996). *Maslach Burnout Inventory manual.* Palo Alto, CA: Consulting Psychologists Press.

Meyer, J. W. (1977). The effects of education as an institution. *American Journal of Sociology, 83,* 55–77. http://dx.doi.org/10.1086/226506

Mulkey, L. M., Catsambis, S., Steelman, L. C., & Crain, R. L. (2005). The long-term effects of ability grouping in mathematics: A national investigation. *Social Psychology of Education, 8,* 137–177. http://dx.doi.org/10.1007/s11218-005-4014-6

Muthén, B. O., & Muthén, L. K. (1998–2013). *Mplus user's guide.* Los Angeles, CA: Author.

Nagengast, B., & Marsh, H. W. (2012). Big fish in little ponds aspire more: Mediation and cross-cultural generalizability of school-average ability effects on self-concept and career aspirations and science. *Journal of Educational Psychology, 104,* 1033–1053. http://dx.doi.org/10.1037/a0027697

Neumann, M., Becker, M., & Maaz, K. (2013). Die abkehr von der traditionellen dreigliedrigkeit im sekundarschulsystem: Auf unterschiedlichen wegen zum gleichen ziel? [The renunciation of the traditional three-tier structure in the secondary school system: On different paths to the same destination?]. *Recht der Jugend und des Bildungswesens, 61,* 274–292.

Oakes, J. (1985). *Keeping track how schools structure inequality.* Binghamton, NY: Vail-Ballou Press.

OECD. (2014). *PISA 2012 technical report.* Paris, France: OECD Publishing.

Pallas, A. M., Entwisle, D. R., Alexander, K. L., & Stluka, M. F. (1994). Ability-grouping effects: Instructional, social, or institutional. *Sociology of Education, 67,* 27–46. http://dx.doi.org/10.2307/2112748

Preckel, F., & Brüll, M. (2010). The benefit of being a big fish in a big pond: Contrast and assimilation effects on academic self-concept. *Learning and Individual Differences, 20,* 522–531. http://dx.doi.org/10.1016/j.lindif.2009.12.007

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hamman, M., Klieme, E., & Pekrun, R. (2007). *PISA 2006. Die Ergebnisse der dritten internationalen vergleichsstudie* [PISA 2006. Results from the third international comparative study]. Münster, Germany: Waxmann.

Protsch, P., & Solga, H. (2015). How employers use signals of cognitive and noncognitive skills at labor market entry: Insights from field experiments. *European Sociological Review, 31,* 521–532. http://dx.doi.org/10.1093/esr/jcv056

Protsch, P., & Solga, H. (2016). The social stratification of the German VET system. *Journal of Education and Work, 29,* 637–661. http://dx.doi.org/10.1080/13639080.2015.1024643

Reuman, D. A. (1989). How social comparison mediates the relation between ability-grouping practices and students' achievement expectancies in mathematics. *Journal of Educational Psychology, 81,* 178–189. http://dx.doi.org/10.1037/0022-0663.81.2.178

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York, NY: Wiley. http://dx.doi.org/10.1002/9780470316696

Schoon, I., McCulloch, A., Joshi, H. E., Wiggins, R. D., & Bynner, J. (2001). Transitions from school to work in a changing social context. *Young, 9,* 4–22. http://dx.doi.org/10.1177/110330880100900102

Schwanzer, A. D., Trautwein, U., Lüdtke, O., & Sydow, H. (2005). Entwicklung eines instruments zur erfassung des selbstkonzepts junger erwachsener [Development of a questionnaire on young adults' self-concept]. *Diagnostica, 51,* 183–194. http://dx.doi.org/10.1026/0012-1924.51.4.183

Schwarzer, R., Lange, B., & Jerusalem, M. (1982). Selbstkonzeptentwicklung nach einem bezugsgruppenwechsel [Self-concept development after a reference-group change]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 14,* 125–140.

Seaton, M., Marsh, H. W., & Craven, R. G. (2009). Earning its place as a pan-human theory: Universality of the big-fish-little-pond effect across 41 culturally and economically diverse countries. *Journal of Educational Psychology, 101,* 403–419. http://dx.doi.org/10.1037/a0013838

Seaton, M., Marsh, H. W., & Craven, R. G. (2010). Big-fish-little-pond effect: Generalizability and moderation—Two sides of the same coin. *American Educational Research Journal, 47,* 390–433. http://dx.doi.org/10.3102/0002831209350493

Shavit, Y., & Müller, W. (2000). Vocational secondary education. Where diversion and where safety net? *European Societies, 2,* 29–50. http://dx.doi.org/10.1080/146166900360710

Solga, H. (2004). Increasing risks of stigmatization: Changes in school-to-work transitions of less-educated West Germans. *Yale Journal of Sociology, 4,* 99–129.

Sung, Y.-T., Huang, L.-Y., Tseng, F.-L., & Chang, K.-E. (2014). The aspects and ability groups in which little fish perform worse than big fish. Examining the big-fish-little-pond effect in the context of school tracking. *Contemporary Educational Psychology, 39,* 220–232. http://dx.doi.org/10.1016/j.cedpsych.2014.05.002

Thijs, J., Verkuyten, M., & Helmond, P. (2010). A further examination of the big-fish–little-pond effect: Perceived position in class, class size, and gender comparisons. *Sociology of Education, 83,* 333–345. http://dx.doi.org/10.1177/0038040710383521

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology, 98,* 788–806. http://dx.doi.org/10.1037/0022-0663.98.4.788

Trautwein, U., Lüdtke, O., Marsh, H. W., & Nagy, G. (2009). Within-school social comparison: How students perceive the standing of their class predicts academic self-concept. *Journal of Educational Psychology, 101,* 853–866. http://dx.doi.org/10.1037/a0016306

Tymms, P. (2001). A test of the big fish in a little pond hypothesis: An investigation into the feelings of seven-year-old pupils in school. *School Effectiveness and School Improvement, 12,* 161–181. http://dx.doi.org/10.1076/sesi.12.2.161.3452

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39,* 111–133. http://dx.doi.org/10.1207/s15326985ep3902_3

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45,* 1–68.

Van Houtte, M., Demanet, J., & Stevens, P. A. J. (2012). Self-esteem of academic and vocational students: Does within-school tracking sharpen the difference? *Acta Sociologica, 55,* 73–89. http://dx.doi.org/10.1177/0001699311431595

Van Houtte, M., & Stevens, P. A. J. (2009). Study Involvement of academic and vocational students: Does between-school tracking sharpen the difference? *American Educational Research Journal, 46,* 943–973. http://dx.doi.org/10.3102/0002831209348789

Warm, T. A. (1989). Weighed likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427–450. http://dx.doi.org/10.1007/BF02294627

Wouters, S., De Fraine, B., Colpin, H., Van Damme, J., & Verschueren, K. (2012). The effect of track changes on the development of academic self-concept in high school: A dynamic test of the big-fish-little-pond effect. *Journal of Educational Psychology, 104,* 793–805. http://dx.doi.org/10.1037/a0027732

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER Conquest version 2.0: Generalized item response modelling software.* Melbourne, Australia: Acer Press.

Zeidner, M., & Schleyer, E. J. (1999). The big-fish-little-pond effect for academic self-concept, test anxiety, and school grades in gifted children. *Contemporary Educational Psychology, 24,* 305–329. http://dx.doi.org/10.1006/ceps.1998.0985

Received March 16, 2016
Revision received November 18, 2016
Accepted November 19, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 6, 871–888

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000156

The Effects of Student Characteristics on Teachers' Judgment Accuracy: Disentangling Ethnicity, Minority Status, and Achievement

Johanna Kaiser
Kiel University

Anna Südkamp
TU Dortmund University

Jens Möller
Kiel University

Teachers' judgments of students' academic achievement are not only affected by the achievement themselves but also by several other characteristics such as ethnicity, gender, and minority status. In real-life classrooms, achievement and further characteristics are often confounded. We disentangled achievement, ethnicity and minority status and examined whether the achievement of ethnic minority students is judged according to the predominant expectation (expectation hypothesis) and whether teachers' judgment accuracy is influenced by students' ethnicity or their minority status (ethnicity hypothesis or minority hypothesis). We conducted 4 experimental studies with a computer simulation (the Simulated Classroom). In Studies 1 and 2 with $N = 34$ and $N = 30$ participants, we implemented Turkish (Study 1) and Asian students (Study 2) as minorities. In contrast to the expectation hypothesis, the expectations attributed to the achievement of ethnic minority students did not bias teachers' judgments. In both studies we found greater judgment accuracy for ethnic minority students, thereby probing the ethnicity hypothesis. In Study 3 with $N = 48$ participants, we further disentangled ethnicity and minority using German students as minority students, thus probing the minority hypothesis. Again, minority students were judged more accurately. Implementing gender (male vs. female) as the minority characteristic in Study 4, with $N = 52$ participants, yielded the same result: Minority students were judged more accurately, therefore supporting the minority hypothesis. Thus, classroom characteristics need to be considered in research on teachers' judgment accuracy to clarify the influence of individual student characteristics and composition effects beyond individual effects.

Keywords: teacher judgment, judgment accuracy, academic achievement, minority students

One of the main tasks that teachers have is judging their students' academic achievement. Acknowledging the importance of teachers' ability to assess students' achievement accurately, the American Federation of Teachers, National Council on Measurement in Education, and National Education Association (1990) have developed standards for teacher competence in the educational assessment of students. In the same vein, the National Board for Professional Teaching Standards (2002) considers the ability to accurately gauge student achievement to be an essential part of the professional knowledge and skills of teachers. Voss, Kunter, and Baumert (2011) also underline the importance of professional knowledge of classroom assessment as an aspect of teaching quality. Teachers often use their implicit or explicit judgments of student achievement as a basis for their instructional decisions (e.g., Alvidrez & Weinstein, 1999; Clark & Peterson, 1986; Hoge, 1983; Hoge & Coladarci, 1989; Voss et al., 2011). Choosing classroom activities, selecting learning materials, defining the difficulty of assigned tasks, establishing the pace of covering content, and so forth depends on assessments concerning students' learning prerequisites. Therefore, teachers need to be able to judge their students accurately, in order to provide effective instruction (Elliott, Lee, & Tollefson, 2001; Ready & Wright, 2011). Furthermore, accurate information on students' academic achievement is crucial for meaningful placement decisions (Helwig, Anderson, & Tindal, 2001). If teachers accurately recognize learning difficulties, students can be provided with suitable interventions.

Essential for a successful classroom assessment is a teacher's ability to judge his or her students accurately (Spinath, 2005; Thiede et al., 2015). However, research showed that teacher judgments are far from perfect and that there is a high variability in the judgment accuracy of teachers (Südkamp, Kaiser, & Möller, 2012). In their heuristic model, Südkamp et al. (2012) defined teacher judgment accuracy as the correspondence between teachers' judgments of students' academic achievement and students' actual achievement as measured by a standardized test. In most

This article was published Online First December 12, 2016.

Johanna Kaiser, Department of Psychology, Kiel University; Anna Südkamp, Faculty of Rehabilitation Sciences, TU Dortmund University; Jens Möller, Department of Psychology, Kiel University.

We thank Christine Schubert for her support during the data collection and Isabelle Kaufmann for her editorial support during the preparation of this article. The research reported in this article is part of the project "Development of Competence in Prospective Teachers" led by Jens Möller, Kiel University, funded by the Federal Ministry of Education and Research (BMBF; promotional reference 01JH0915).

Correspondence concerning this article should be addressed to Johanna Kaiser, Department of Psychology, Kiel University, Olshausenstraße 75, 24118 Kiel, Germany. E-mail: kaiser@psychologie.uni-kiel.de

studies, the correlation between the two is used as a measure of this correspondence. However, other indicators, such as the average difference between teacher judgments and students' actual performance, can also be used. As Spinath (2005) has demonstrated, distinguishing between indicators of teacher judgment accuracy is crucial as they are typically correlated close to zero. Südkamp et al. (2012) further described that student characteristics are thought to influence judgment accuracy. One student characteristic that is prone to substantially bias teachers' judgments is ethnicity (Tenenbaum & Ruck, 2007). In the United States, as well as in European countries, the situation is comparable: Compared with Caucasians, a lot of ethnic minority students score lower on large scale assessments and are more likely to drop out of school (Luciak, 2003, 2006). The generally low achievement of ethnic minority students seems to lead to low teacher expectations (e.g., Tenenbaum & Ruck, 2007). But, it is not yet clear whether, and if so to what extent, these expectations affect teacher judgment and accordingly teacher judgment accuracy. Furthermore, as low achievement and ethnic minority status are often confounded variables it is difficult to unravel effects of achievement and effects of ethnicity and minority status in real classrooms. Therefore, the central purpose of our present studies was to disentangle ethnicity, minority status and achievement and to analyze whether the ethnicity and/or the minority status of students affected the judgment and the judgment accuracy of teachers in an experimental setting. With regard to the heuristic model by Südkamp et al. (2012), we analyzed the unconfounded effects of central student characteristics with our experimental approach.

Teacher Judgment Accuracy

Current research on teacher judgment accuracy has predominantly focused on student achievement, comparing students' performances on a standardized achievement test with teachers' judgments of students' achievement. In a meta-analysis with 75 studies, Südkamp et al. (2012) found a mean effect size of $Zr = .63$ for the relationship between teachers' judgments and students' achievement. Ranging from $r = -.03$ to $r = .84$, the variability of correlations between studies was very high, suggesting large differences in teachers' abilities to judge students' achievement accurately. These findings confirm the conclusions of the review of comparative studies conducted by Hoge and Coladarci (1989) that report a median of correlations of $r = .66$ (range between $r = .28$ and $r = .92$) between teachers' judgments' and students' achievement on a standardized achievement test. In their heuristic model Südkamp et al. (2012) proposed four factors that influence teacher judgment accuracy: judgment characteristics (e.g., indirect vs. direct judgment), test characteristics (e.g., subject area covered), teacher characteristics (e.g., professional experience, cognitive abilities), and student characteristics (e.g., achievement level, ethnicity, minority status). Whereas information about judgment characteristics and test characteristics could be considered in the meta-analysis by Südkamp et al. (2012), too few studies reported comparable information on teacher characteristics and student characteristics and their correspondence to teachers' judgment accuracy. In contrast to teacher characteristics, student characteristics such as gender, ethnicity, or social status are discussed more often and more extensively as moderators of teacher judgment accuracy. This might be because a bias in teachers' judgments caused by particular student characteristics may lead to differential treatment of students—like fewer response opportunities or less positive feedback—and thus to discrimination (Ready & Wright, 2011). For instance, Bennett, Gottesman, Rock, and Cerullo (1993) studied the effect of students' behavior on teachers' judgments of students' academic skills (e.g., word recognition, arithmetic), controlling for students' gender and academic achievement in a path model. Teacher judgment accuracy, as measured by the direct effect of teachers' judgments on students' academic achievement, ranged between $ß = .29$ and $.61$. Teachers judged students with bad behavior as academically poorer in comparison with students who behaved satisfactorily, regardless of students' actual academic achievement. Hurwitz, Elliott, and Braden (2007) examined the relationship between students' disability status (learning disability, speech/language impairment, emotional-behavioral disability, orthopedic/other health impairment) and teachers' judgment accuracy. They reported a greater item-level percentage agreement between teachers' judgments and students' test scores for students without disabilities than for students with disabilities. As can be seen by these findings, individual student characteristics may have an impact on teachers' judgment accuracy. The question of how student ethnicity affects teacher judgment accuracy is often discussed but has not yet been satisfactorily answered.

Student Ethnicity and Teacher Judgments

Studies on the effects of students' ethnicity on teachers' judgments are heterogeneous in terms of their theoretical and methodological approaches. In our review of literature, we carefully reflect the differences, in order to contribute to a better understanding of the current state of research. There is a large body of research that has examined differences in teachers' expectations toward students of different ethnicities. Many of these studies were inspired by the famous Pygmalion study (Rosenthal & Jacobson, 1968), in which some students within a classroom were randomly labeled as "late bloomers." The promising label resulted in a greater increase of students' academic achievement and IQ scores in comparison to their nonlabeled peers. Rosenthal and Jacobson (1968) called higher teacher expectations toward students labeled as late bloomers to be responsible for this effect.

In the same vein, the question of whether a student's ethnicity plays a role in the expectation that teachers have is considered particularly relevant as teachers may treat students differently based on their expectations and might discriminate against ethnic minority students (McKown & Weinstein, 2008; Tenenbaum & Ruck, 2007). In their meta-analysis, Dusek and Joseph (1983) found lower expectations of students' achievement for African American students and Mexican students compared with European American students (see also Baron, Tom, & Cooper, 1985). Tenenbaum and Ruck (2007) corroborated these findings in their more recent meta-analysis. They found that teachers held the highest expectations for Asian American students compared with all other student groups, and the expectations for European American students were higher than the expectations for Latino/a and African American students. In the Netherlands, van den Bergh, Denessen, Hornstra, Voeten, and Holland (2010) found lower expectations for students of particular ethnic minorities (students of Turkish or Moroccan background). Research on teacher expectations has used experimental and nonexperimental designs. In

their meta-analysis on teachers' expectations toward racial minorities, Tenenbaum and Ruck (2007) found that participants provided more positive ratings for European American children than for children of other ethnicities when they viewed a photograph, watched a videotape, or listened to an audiotape, rated their own students or rated stimulus students. In contrast, participants provided more negative ratings toward European American children when they rated vignettes. Teacher expectation research is also characterized by a variety of approaches of measuring teachers' judgments. As an example, van den Bergh et al. (2010) collected teachers' ratings of students' current skills based on the teachers' observations ("He or she performs well in school") as well as teachers' ratings of students' future performance ("He or she will probably have a good school report at the end of this school year"). In turn, Woodworth and Salzer (1971) asked teachers to complete a teacher rating sheet after they had listened to Black and White children's recorded social study reports.

Concerning studies of teacher expectations, there is criticism that objective measures of student achievement were rarely employed (Jussim & Harber, 2005; Ready & Wright, 2011). Therefore, it cannot be clarified if differences in teacher expectations for different groups of students may in fact stem from actual differences between these students. Jussim and Harber (2005) explicitly asked for more studies dealing with the accuracy of teachers' judgments associated with students from differing social and demographic groups, thus comparing teachers' judgments to students' actual achievement on a standardized test.

Addressing these implications, Ready and Wright (2011) used nationally representative data and three-level hierarchical linear models to explore the relationship between teachers' judgments and kindergarten children's sociodemographic backgrounds. Whereas half of the differences in teachers' judgments could be attributed to actual achievement differences between the students, the authors found substantial differences in teachers' judgments across student ethnic subgroups that cannot be explained by actual differences between these groups. Evidently, the literacy skills of Asian children, not speaking English as their primary home language, Hispanic, and African American children were underestimated compared with those of European American children. In contrast to Ready and Wright (2011), other studies showed no differences in teachers' judgment accuracy for students of different ethnicities. Recently, Peterson, Rubie-Davies, Osborne, and Sibley (2016) found no main effect for student ethnicity on teachers' explicit academic judgments in an ANCOVA controlling for beginning-of-year achievement scores. Baker, Tichovolsky, Kupersmidt, Voegler-Lee, and Arnold (2015) studied teachers' judgments of preschoolers' academic skills using hierarchical linear modeling. They found no association between students' ethnicity and teachers' ratings. Using multiple regression analysis, Madon et al. (1998) found that teachers did not rely on ethnic stereotypes when judging students' performance, talent, or effort (ß-coefficients $< .01$). Interpreting their findings, Madon et al. (1998) state that teachers judged students on the basis of their achievement and motivation and teachers' judgments were mostly accurate. Hachfeld, Anders, Schroeder, Stanat, and Kunter (2010) examined teachers' judgment accuracy in a sample of $N = 305$ mathematics teachers. Prior to the analyses, they coded teachers' judgment accuracy in terms of an over- or underestimation of students' performance and used this variable as the dependent measure in a multilevel analysis. Students' ethnic background did not have an effect on teacher judgment accuracy. To further illustrate the unclear evidence, in New Zealand, Rubie-Davies, Hattie, and Hamilton (2006) found that teachers' judgments were accurate for New Zealand European students as well as for Maori and Asian students. Analyses of variance with teachers' judgments of students' academic achievement as the dependent variable revealed that only the judgments for Pacific Island students were higher than their actual achievement. In summary, the evidence on teacher judgment accuracy for students of different ethnicities is ambiguous: Whereas some studies found no differences in teachers' judgment accuracy across students of different ethnicities (Hachfeld, Anders, Schroeder, Stanat, & Kunter, 2010; Madon et al., 1998), others found systematic inaccuracy in terms of over- or underestimation for students of different ethnic backgrounds (Ready & Wright, 2011; Rubie-Davies et al., 2006). Concerning different measures of teacher judgment accuracy, the abovementioned studies predominantly used the average difference between teachers' judgments and students' actual performance. In our study, we consider both—the average difference as well as the correlation between teachers' judgments and students' actual performance. In order to complement prior research, we also take the classroom composition into account.

Classroom Composition Effects on Teacher Judgments

In research on teachers' judgment processes, it is reasonable to consider class characteristics in addition to individual student characteristics (Westphal et al., in press). For instance, this is illustrated by the big-fish-little-pond-effect (BFLPE; Marsh, 1987), the phenomenon that the individual achievement of students positively impacts teachers' judgments and students' self-concepts, whereas the class level negatively impacts teachers' judgments, thereby disadvantaging students who are in high-achieving classes (e.g., Trautwein & Baeriswyl, 2007; Trautwein, Lüdtke, Köller, & Baumert, 2006). Südkamp and Möller (2009) were able to replicate this reference group effect in an experimental setting with a simulated classroom scenario. Students with identical achievement were graded less favorably in a class with a high average level of achievement compared with a class with a lower average level of achievement.

Apart from the class achievement level, student characteristics like ethnicity and gender were also considered in the literature on classroom composition effects (Hattie, 2002; van Ewijk & Sleegers, 2010). Usually, such characteristics are not distributed equally but constitute minorities and majorities in the classroom that also need to be considered, as shown by Ready and Wright (2011). The authors found a positive association between the social status of students and teachers' judgments: Higher social status led to higher teacher judgments while controlling for individual student achievement. Interestingly, this association was stronger at the class level: Students' socioeconomic status (SES) influenced teachers' judgments more in low-SES classes than in high-SES classes. For example, this means that a teacher assesses a socioeconomically disadvantaged child's achievement lower in a high-SES classroom than a child with the same SES and achievement in a low-SES classroom. Ready and Wright (2011) explained this effect by the minority members' salience. In a class mainly consisting of socioeconomically disadvantaged children, a high-SES

child was more likely to stand out than a low-SES child (Ready & Wright, 2011). Summarizing their results, the authors underlined the importance of considering the class composition when examining teacher judgment accuracy. Classes differ in many aspects, like their achievement level or their homogeneity, which could lead to some students being more salient than others and thus attracting more attention. When minorities attract attention, it is plausible to expect better information processing for these salient stimuli and more accurate judgments. It could be assumed that information processing for salient stimuli acts as a mediator between increased attention and more accurate judgments. Salience seems to increase organization and consistency of information (Fiske & Taylor, 1991). The information is "stored at the top of the mental heap" (Taylor & Fiske, 1978, p. 270) and therefore easily available. This needs to be taken into account for research on the impact of ethnicity on teacher judgment accuracy. On the one hand, a student's ethnicity may have an impact on the teacher's judgment accuracy. On the other hand, the distribution of ethnicities within a class may also make a difference: A student with a certain ethnicity may either be a member of a minority group or be a member of a majority group and therefore be more or less salient. Thus, research on judgment accuracy needs to be able to disentangle ethnicity and minority status. If there are differences in teacher judgment accuracy for students of different ethnicities it needs to be made clear whether these differences occur because of the students' ethnicity or because of a classroom composition effect where a particular ethnicity stands out.

Present Research

Based on previous findings, the question of whether, and if so, how, teacher judgment accuracy is influenced by the ethnicity of students remains without a clear answer. Research on teacher judgment accuracy toward particular student characteristics, like ethnicity, has to take the classroom composition into account (McKown & Weinstein, 2008). Students with a migration background often form a minority in a class, so that ethnicity and minority status are confounded. Lesser or greater teacher judgment accuracy for the members of an ethnic minority group can be attributed to their ethnicity or to their minority status. The matter gets more complicated as ethnic minority students often show lower achievement than ethnic majority students. Teachers have more difficulties in judging low-achieving students correctly compared with high-achieving students (Begeny, Eckert, Montarello, & Storie, 2008; Feinberg & Shapiro, 2009). Research on teacher judgment accuracy in relation to student characteristics, like ethnicity, thus has to disentangle students' ethnicity, minority status and achievement, which is not easy for field studies because these student characteristics are confounded. Therefore, we took an experimental approach in the present study. We conducted four studies in differently composed classes addressing the research questions within a simulated classroom. In a simulated classroom we can experimentally manipulate students' minority status, ethnicity, and achievement independent of each other and control for further influences. We examined teachers' judgment accuracy for different (ethnic) minority students. In Study 1, students with a Turkish migration background served as a minority in Germany. In Study 2, students with an Asian migration background formed the minority. In Study 3, we used German students as a minority (and Turkish students as the majority) to disentangle ethnicity and minority status. Finally, in Study 4, we chose students' gender as a minority characteristic to test whether it is the minority status or the ethnicity that really affects teacher judgment accuracy.

Our research aim was to examine whether teachers' judgments and teachers' judgment accuracy were influenced by students' ethnicity and their minority status. First, we wanted to analyze whether the achievement of students from different ethnic backgrounds was judged according to the predominant expectations (Research Question 1). Therefore, we compared the average level of teacher judgments for students of different ethnic backgrounds. We expected lower teacher judgments of students' academic achievement for an ethnicity for whom low achievement is generally expected (Turkish students) in comparison with teacher judgments on other students, and we expected higher judgments for students for whom high achievement is generally expected (Asian students) compared with others—while controlling for actual achievement (*expectation hypothesis*). Second, we wanted to examine whether students' ethnicity or their minority status moderated teachers' judgment accuracy (Research Question 2). We operationalized judgment accuracy as the relationship between students' actual achievement and teachers' judgments. In the course of our four studies, two hypotheses were tested against each other concerning the second research question: The *ethnicity hypothesis*, which postulates that judgment accuracy differs for judgments on students from whom low or high academic achievement is generally expected compared with other students in a class, and the *minority hypothesis*, which postulates that not the ethnicity but the minority status as such (independent of the ethnicity) leads to different teacher judgment accuracy.

Apart from ethnicity and gender we did not vary further student characteristics that could form minority/majority proportions and we distributed high and low achievement equally across minority and majority students. We took ethnicity and gender as student characteristics as they should not be relevant for a student's achievement per se. It was beyond the scope of our studies to examine students who are in the minority because of their low (or high) achievement.

Studying Teacher Judgment Accuracy in an Experimental Setting: The Simulated Classroom

To investigate these questions, systematic manipulations of student characteristics are necessary. Thus, to study the influence of students' ethnicity on teachers' judgment accuracy with high internal validity, a possible confounding between students' achievement and further student characteristics needs to be disentangled. An experimental setting provides ideal conditions to study the influence of student characteristics while controlling for test characteristics and judgment characteristics.

The Simulated Classroom (see also Fiedler, Freytag, & Unkelbach, 2007; Fiedler, Walther, Freytag, & Plessner, 2002; Kaiser, Retelsdorf, Südkamp, & Möller, 2013) is a computer simulation of an instructional situation in which student factors, for example, achievement (in terms of the proportion of correct answers), engagement (in terms of participation in class), gender (as indicated by a photograph and/or name), and instructional factors (subject, number of students, lesson length, content covered) can be experimentally manipulated. The participant takes over the role

of a teacher interacting with simulated students. In doing so, the participant directs questions at the students and observes their responses. Later, the participant's task is to assess students' achievement. The Simulated Classroom is programmed in Java and participants work individually on personal computers. To begin, participants are given an introduction to the Simulated Classroom via instruction video. Before the lesson starts, they are informed about the students' grade and the topic of the lesson. Their task is to select topic-specific questions from a menu of possible questions and to address these questions to the students in their "class." In order to get familiar with the functionality of the Simulated Classroom, participants complete a short tutorial sequence before the actual lesson. The students are represented by a photograph and their name written on virtual desks (see Figure 1). The names are randomly chosen from the most popular children's names in the year the simulated student would theoretically have been born. Photographs, names (and thus gender), and seating positions are allocated at random (making sure that the gender and name allocated matched the photo). This randomization leads to a different class for each participant, but it has the advantage that photographs, names, seating positions, and achievement parameters are not confounded. Each question selected by the "teacher" is displayed at the bottom left of the computer screen. Some of the students then volunteer to answer the question in accordance with their predefined engagement parameters. These students are indicated by the coloring of their desks, which changes from black to yellow. The teacher may call on the students by a mouse click on the respective desk. That student then gives a correct or an incorrect answer depending on his or her predefined ability parameter. The answer is displayed at the bottom right of the screen. In order to control for an influence of participants' content knowledge or his or her pedagogical content knowledge (see Shulman, 1987), the correctness or the incorrectness of the answer is clearly marked. If it is correct, it appears in a green box. Otherwise, one of three possible incorrect answers appears in a red box. This variation of incorrect answers reduces the probability that the same incorrect answer will be given consecutively by different students.

Once a question-and-answer sequence has been completed, the teacher can direct either the same question or a new question to the students. The experimenter can vary the length of the lesson and within the given timeframe the teacher can ask any number of questions. The proportion of correct answers provided by each student is experimentally varied and represents the level of student achievement. At the end of the "lesson," participants are asked to

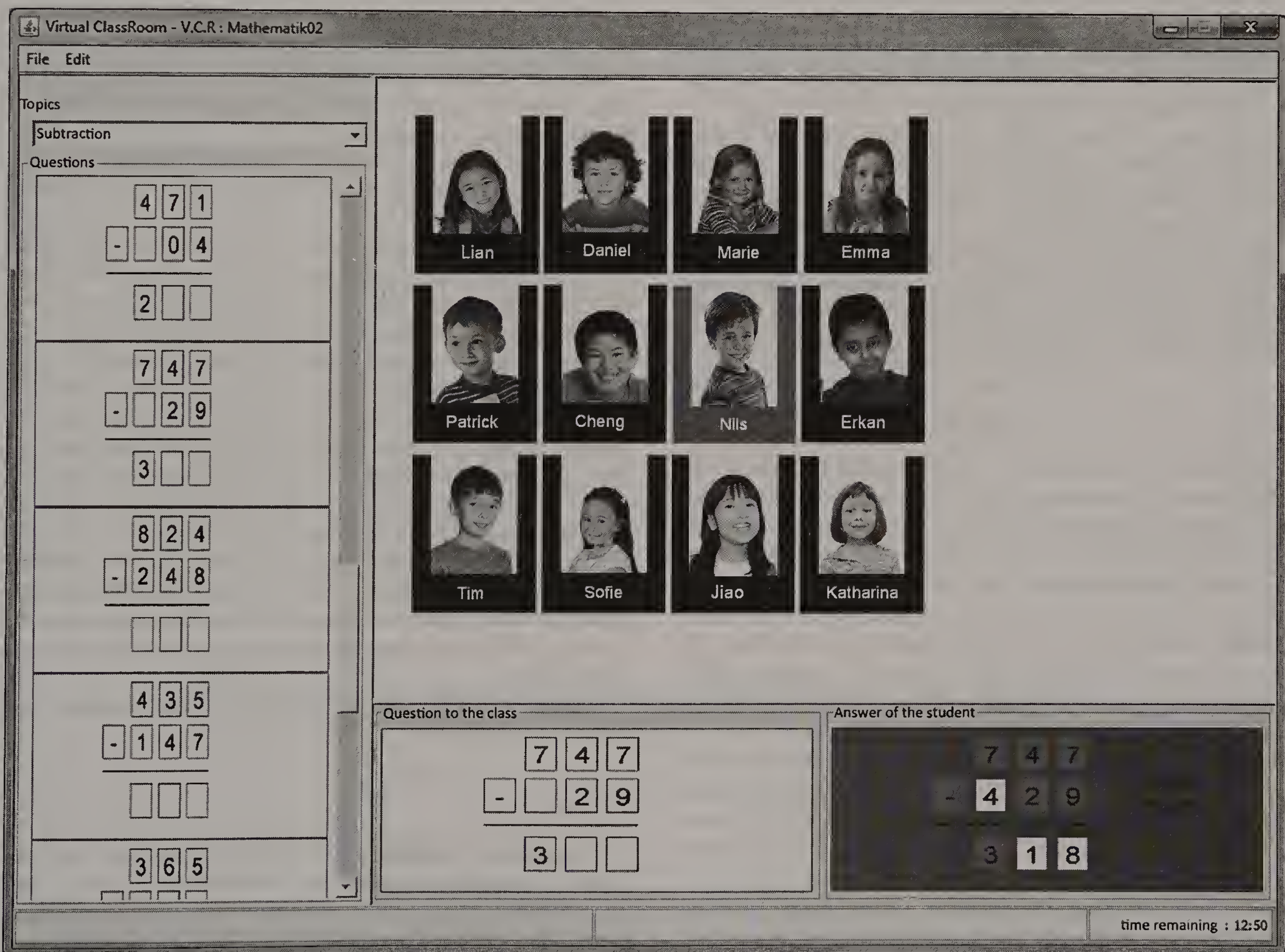


Figure 1. Screenshot of the Simulated Classroom.

judge the proportion of correct answers given by each student. Hence, the teachers give an informed judgment according to Südkamp et al. (2012). Based on this, it is possible to gauge the extent to which the participants' judgments correspond with the students' actual achievement. The congruence between teachers' ratings and the students' achievement measure is given.

For the following studies, we simulated a third grade mathematics class. Participants had 17 min to select questions from the topic areas of the number line, addition, subtraction, multiplication, measurement units, or word problems. These tasks were taken from the German Mathematics Test for Third Graders (DEMAT 3+; Roick, Gölitz, & Hasselhorn, 2004). We chose the questions that had a medium difficulty and to provide a sufficiently large pool of questions, we developed a number of additional questions that were constructed according to the same principles as the DEMAT 3+ tasks. To avoid effects of task difficulty, we told participants that all tasks were of approximately the same (moderate) difficulty level. With this operationalization, the participants' judgment accuracy did not depend on their content knowledge or their pedagogical content knowledge. The engagement parameter was set to a 50% probability of volunteering to answer a question. Thus, it could be ensured that all students had the same chance to be called on. Students could only be called on when they volunteered to answer a question. To answer our research question, we conducted four studies with the Simulated Classroom, in which we experimentally manipulated the class composition concerning students' ethnicity and their gender.

The Simulated Classroom has already proven to be a valuable tool in educational psychology to answer questions regarding teacher judgment accuracy (see Kaiser, Helm, Retelsdorf, Südkamp, & Möller, 2012; Kaiser et al., 2013; Südkamp & Möller, 2009; Südkamp, Möller, & Pohlmann, 2008). For example, the effect of mutual influences of students' achievement and engagement on teachers' judgments often found in the field could be replicated with the Simulated Classroom (Kaiser et al., 2013).

Study 1

In Study 1, we investigated whether students belonging to an ethnic minority, from whom low academic achievement is usually expected, are judged lower than students of German origin. In order to do so, students' achievement and ethnicity was disentangled and systematically varied in the Simulated Classroom. In Germany, where the studies were conducted, the expectations for students with a Turkish migration background are rather low. The results of the PISA Study 2006 showed that students with a migration background in Germany, and particularly students with a Turkish migration background, on average, achieved only low competence levels (Prenzel et al., 2008). Therefore, we expected that the achievement of students with a Turkish migration background would be judged lower than students of German origin (Research Question 1; *expectation hypothesis*). As for the second research question concerning differential judgment accuracy for ethnic minority students, we tested the ethnicity hypothesis. With the design at hand, the minority hypothesis could not be tested, as we did not disentangle ethnicity and minority (the *minority hypothesis* will be tested in Studies 3 and 4).

Sample

Study 1 drew on data obtained from $N = 34$ elementary school teachers (88.2% female). Thirty-three participants were born in Germany and all of them spoke German during their childhood. Two participants were brought up bilingual. At the time of the data collection, all participating teachers stated that they spoke German at home (one participant spoke German and one additional language at home). The participants' mean age was $M = 45.3$ ($SD = 11.8$) years and on average they had $M = 17.9$ ($SD = 12.4$) years of teaching experience.

Measures

Students' ethnicity. We simulated 12 students (six girls and six boys) in a third grade mathematics class. The names (e.g., Lena) and photographs of eight students suggested that they were of German origin (in the following referred to as the German students); three students' names (e.g., Nesrin) and photographs suggested that the students had a Turkish migration background (in the following referred to as the Turkish students); and the name (e.g., Cheng) and photograph of one student suggested that the student was of Asian origin (in the following referred to as the Asian student). The purpose of having one Asian student in the class was to prevent participants from noticing only two ethnicities and guessing the study's hypotheses.

Students' achievement. To operationalize students' achievement, three achievement levels were chosen: Four students showed low achievement (probability of a correct answer: approx. 20%); four students showed medium achievement (probability of a correct answer: approx. 50%); and four students showed high achievement (probability of a correct answer: approx. 80%). These levels must be considered as parameters. The simulated students' actual achievement in the Simulated Classroom is determined by a probability algorithm, such that the proportion of correct or incorrect answers given by each student corresponds approximately with his or her achievement parameter, depending on the number of questions one student is called on to answer. There was one Turkish student on each achievement level. The Asian student was programmed to show medium achievement. Of the German students, three were on the low level, two were on the medium level, and three were programmed to show high achievement. The achievement was measured on a scale from 0%–100%.

Teachers' judgments. Participants were asked to rate the percentage of correct answers given by each student on a scale from 0%–100%. Therefore, the judgments of the teachers and the students' achievement were on the same scale and could be compared directly.

Number of calls. We recorded the number of times each student was called on.

Results

The statistical analyses for all studies to be reported were identical and therefore presented in more detail here. Table 1 gives an overview of the measures for the different student groups concerning ethnicity (German, Turkish, Asian) and achievement level (low, medium, and high). For each group, the mean of the students' actual achievement level, the mean of teachers' judg-

Table 1
Measures in Study 1

Ethnicity	German (majority)			Turkish (minority)			Asian
Achievement level	Low (n = 102) *M (SD)*	Medium (n = 68) *M (SD)*	High (n = 102) *M (SD)*	Low (n = 34) *M (SD)*	Medium (n = 34) *M (SD)*	High (n = 34) *M (SD)*	Medium (n = 34) *M (SD)*
Actual achievement in percent	18.99 (9.34)	54.55 (13.05)	85.69 (10.31)	19.44 (7.75)	49.07 (9.37)	86.73 (8.26)	50.81 (11.13)
Achievement judgment in percent	37.01 (22.40)	56.15 (18.01)	73.58 (20.49)	27.26 (21.50)	50.94 (18.40)	77.21 (19.58)	56.79 (17.13)
Number of calls	14.39 (7.45)	13.10 (6.99)	12.99 (7.95)	14.94 (6.80)	14.47 (8.94)	13.59 (5.38)	12.53 (6.01)

Note. The n indicates the number of available judgments for different types of students, categorized by ethnicity and achievement level. Note that students' actual achievement levels slightly differ from the predetermined achievement level (20%, 50%, 80%) as a probability algorithm was implemented.

ments of students' achievement, and the average number of calls are reported.

In order to answer the research questions, we used the Mplus software package (Muthén & Muthén, 2010). As each teacher was dealing with several students and provided judgments for them, the data had a hierarchical structure, which may have led to an underestimation of standard errors and thus to biased significance tests if not taken into account (e.g., Hox, 2002). To obtain unbiased estimates, we used the "type = complex" option implemented in Mplus, together with a robust maximum-likelihood estimator (RML; Muthén & Satorra, 1995). Due to the data collection via computer, there were no missing values as participants could only proceed when all assessments were completed. We conducted moderation analyses. In all analyses, teachers' judgments were used as the criterion and were predicted by students' actual achievement as a first predictor and both were on the percentage scale from 0 to 100. Students' actual achievement was centered. To examine whether judgments were biased according to the dominant expectation for Turkish students, the ethnicity (dummy-coded, 0 = German, 1 = Turkish) was included as a further predictor (Research Question 1; *expectation hypothesis*). Thus, the indicator to answer this first research question was whether there is a difference in the mean level of judgments for Turkish students compared with German students (a coefficient for the dummy-variable minority status that is statistically significantly different from zero). We controlled for the actual achievement of the students as it was also included in the analysis.

In order to find a difference in judgment accuracy for Turkish students compared with German students, the interaction of the ethnicity and students' achievement was also included. In this way, we were able to examine whether ethnicity moderates the relationship between teachers' judgments and students' achievement (Research Question 2; *ethnicity hypothesis*). We also added the (previously centered) number of times a student was called on (hereinafter referred to as the number of calls) and the interaction term of the number of calls with the actual achievement as further predictors. With these predictors, it could be examined whether more frequent observations of student achievement led to a different judgment (indicated by a statistically significant coefficient for the predictor number of calls) and whether more frequent observations led to more accurate judgments (a moderating effect of the number of calls on the relationship between teachers' judgments and students' achievement, indicated by a statistically significant coefficient for the interaction term).

For the calculation of standardized regression coefficients, actual achievement, judgments, and the number of calls were z-standardized (M = 0, SD = 1) prior to the analysis. The prior standardization was necessary because the standardization of the product term (as displayed in the standardized solution of the output of the regression analysis) differs from the product term of the standardized variables. The procedure described ensures that it is justified to interpret the coefficients of the output in terms of standardized regression coefficients (Friedrich, 1982). Only Turkish and German students were included in the calculations. The results of the analysis can be seen in Table 2.

The intercept of the unstandardized results showed the judgment for a mean achievement of a German student with a mean number of calls. The coefficient of the first predictor could be interpreted to describe the relationship between students' achievement and teachers' judgments for the German students, thus teacher judgment accuracy for them. There was a considerable relationship between teachers' judgments and students' achievement as indi-

Table 2
Unstandardized and Standardized Results of the Moderation Analysis to Predict Judgments of Student Achievement (Given on the Percentage Scale) in Study 1

Predictor	*B*	*SE*	β	*SE*
Intercept	55.39***	1.37		
Achievement (percentage scale)	.56***	.06	.63***	.07
Ethnicity (0 = German/majority, 1 = Turkish/minority)	−2.97	2.32	−.11	.09
Interaction Ethnicity × Achievement	.18**	.07	.20**	.08
Number of calls	.04	.16	.01	.05
Interaction Number of Calls × Achievement	.00	.01	.02	.05

Note. R^2 = .475.
* $p < .05$. ** $p < .01$. *** $p < .001$.

cated by the statistically significant coefficient for the predictor achievement. The unstandardized coefficient described the increase (0.56% in this case) in teachers' judgments for a 1% increase in students' achievement. The standardized coefficient can be interpreted comparably with correlation coefficients in other studies on teacher judgment accuracy. The unstandardized coefficient of the predictor for the dummy-coded variable showed whether the overall level of teachers' judgments was different for Turkish and German students, controlling for students' achievement (Research Question 1; *expectation hypothesis*). As the statistically nonsignificant coefficient for students' ethnicity showed, we did not find a particular over- or underestimation of the achievement of Turkish students compared to German students (*expectation hypothesis*). The coefficient of the interaction hinted at different relationships between students' achievement and teacher judgments for ethnic minority and majority students (Research Question 2; *ethnicity hypothesis*). We found an ethnicity accuracy effect in the interaction: Students' ethnicity moderated the relationship between teachers' judgments and students' achievement. The unstandardized coefficient described the additional (0.18%) increase (in addition to the increase caused by a one percent increase in students' achievement) in teachers' judgments when a minority student was judged compared with a majority student (0.56% plus 0.18% = 0.74%). The standardized coefficient could be interpreted in the way that it shows the additional effect to the measure of correspondence between students' achievement and teachers' judgments for Turkish students. Turkish students were judged more accurately than German students. The number of calls and the interaction of the number of calls and actual achievement were not statistically significant. Hence, the number of times a student was called on did not moderate the relationship between teachers' judgments and students' achievement.

We preferred the moderation analysis over other measures of accuracy as a strategy of analysis in order to be able to account for the hierarchical data structure and also to be able to analyze moderating effects. Leucht, Tiffin-Richards, Vock, Pant, and Köller (2012) chose a similar analysis.

Discussion

In Study 1, we examined teachers' judgment accuracy concerning students' ethnicity. In our experimental setting we simulated a class consisting mainly of German students and some Turkish students. We experimentally varied students' achievement, thereby disentangling students' ethnicity and achievement.

Contrary to the expectation hypothesis, one main result of Study 1 is that Turkish students' achievement was not judged lower than German students' achievement (Research Question 1). There were no differences in teachers' judgments for Turkish and German students (while controlling for their actual achievement). The moderation analysis supported the ethnicity hypothesis: Turkish students—who formed the minority in Study 1—were judged more accurately than German students. Does this mean that high-achieving minority students are more likely to get the credit for their achievement compared to majority students who are also high-achieving? Otherwise, are low-achieving minority students disadvantaged because their achievement is judged more harshly? To further examine our research questions we chose another student ethnicity to form a minority group in Study 2. Here, we installed an ethnic minority for which high achievement is usually expected. In this way, we were able to rule out that there was a bias in teachers' judgments due to socially desirable judgments.

Study 2

In Study 2, we wanted to answer the question of whether students for whom there are generally positive academic achievement expectations are judged differently than German students. Therefore, we chose Asian students as a minority in a simulated class. Students with an Asian migration background are seen as a "model minority." They are perceived as high-achieving, hardworking, quiet, organized, well-behaved, and respectful to teachers (Chang & Sue, 2003; Schneider & Lee, 1990). In German-speaking countries the expectations for Asian students are also generally high (Weiss, 2000; Weiss, 2007).

The experimental setup and the measures for Study 2 were identical to Study 1. The only difference was that the Asian students composed the minority group. Again, students were supposed to be "instructed" in a third grade mathematics lesson. The analyses were carried out analogously to Study 1. Although we did not find a confirmation for our expectation hypothesis in Study 1, we maintained it in Study 2. We expected Asian students to be judged to have higher achievement than German students. Concerning Research Question 2, we expected to find more accurate judgments for the Asian students (ethnicity hypothesis).

Sample

The initial sample consisted of 31 participants. One participant had to be excluded from the analyses because the participant's data indicated no serious commitment to the task. Accordingly, the sample comprised N = 30 elementary school teachers (100% female). Twenty-eight of the participants were born in Germany, all of them spoke German in their parental home and two of them were brought up bilingual. All participating teachers spoke German at home at the time of the data collection. On average, the participants were M = 41.4 (SD = 11.8) years old and had M = 14.5 (SD = 10.4) years of teaching experience.

Measures

Students' ethnicity. There were eight German students, three Asian students and one Turkish student in the Simulated Classroom.

Students' achievement. As in Study 1, we chose three achievement levels (probability of a correct answer: approx. 20%, 50%, and 80%). The distribution of students' ethnicity across the achievement levels did not differ from Study 1—there was one Asian student on each achievement level, the Turkish student showed medium achievement, three German students were on each of the high and low achievement levels and two were on the medium achievement level.

Teachers' judgments. The judgments of students' achievement were again given on a scale from 0%–100%.

Number of calls. We recorded the number of calls for each student.

Results

All statistical analyses were carried out analogously to Study 1. Table 3 presents means and standard deviations for the measures of the different student groups.

As in Study 1, we ran a moderation analysis in order to answer the question of whether teachers' judgments were biased toward ethnic minority students. Asian and German students were taken into account for the calculations, the Turkish student was excluded. Teachers' judgments were predicted by five predictors: students' actual achievement; students' ethnicity (dummy-coded, 0 = German, 1 = Asian); the interaction term of actual achievement with ethnicity; the number of calls; and the interaction of number of calls with actual achievement (see Table 4).

As can be seen in Table 4, actual achievement served as a strong predictor of teachers' judgments. Asian students were not judged higher than German students, but again, students' ethnicity moderated the relationship between actual achievement and judgments—Asian students were judged more accurately, indicating an ethnicity accuracy effect. Neither the number of calls nor its interaction with actual achievement were statistically significant predictors for teachers' judgments.

Discussion

In Study 2, we tested for differences in teachers' judgments and teachers' judgment accuracy for Asian students, as an ethnic minority, and German students, as a majority, in a simulated class.

The moderation analysis that was conducted in order to clarify whether Asian students were judged to have higher achievement than German students did not support the expectation hypothesis. Asian students were not judged more favorably than German students. Our expectation hypothesis was not supported. However, we again found evidence for the ethnicity hypothesis: Asian students were judged more accurately, regardless of their achievement being high or low. In this aspect, the results of Study 2 replicated those of Study 1.

In the previous Studies 1 and 2, we did not find biased judgments toward ethnic minority students (Research Question 1) but found more accurate judgments for ethnic minority students (Research Question 2). In Study 1, simulated Turkish students constituted the minority, and in Study 2, simulated Asian students constituted the minority in the Simulated Classroom. As the findings of both studies did not meet the expectation hypothesis of Research Question 1, we rejected it. However, for Research Question 2, we found evidence for the ethnicity hypothesis, assuming greater judgment accuracy for ethnic minority students.

A possible explanation for these findings includes the speculation, that we managed to disentangle achievement and ethnicity but still confounded ethnicity and minority. In Studies 1 and 2, Turkish and Asian students served as minority groups in their respective classes compared with German students. We concentrated on the effects of individual student characteristics on teachers' judgments but did not focus on the class composition. Therefore, we further disentangled the confounded variables of minority and ethnicity in Study 3.

Study 3

In Study 3, we examined whether belonging to a minority, independent from ethnicity, can account for the differences in judgment accuracy (minority hypothesis). Therefore, we simulated a class with a minority of German students and a majority of Turkish students. Grade, subject, available questions, and further settings in the Simulated Classroom were the same as in Studies 1 and 2. Based on the findings in Study 1 and 2, we expected to find greater judgment accuracy for German students, being the minority. Thus, in Study 3 we tested the minority hypothesis, as it was the first study in our progression of studies where this hypothesis could be examined.

Sample

The sample comprised N = 48 educational science students (62.5% female); 29 of them were training to teach secondary school (60.5%) and 19 were pursuing a degree in pedagogics (39.5%). The participants of the study were students from a German university. They were studying in various subject combinations and were required to study at least two subjects: 91.8% were studying one humanities and social sciences subject, 60.4% were studying at least one language, 41.7% were studying at least one science or mathematics subject, and 18.8% were studying one other subject. On average, they were in their fourth semester of study (M = 4.4, SD = 3.4) and were M = 24.1 (SD = 3.8) years old.

Measures

Students' minority status. We simulated a class consisting of 10 students: eight Turkish students (majority group) and two German students (minority group).

Table 3
Measures in Study 2

Ethnicity	German (majority)			Asian (minority)			Turkish
Achievement level	Low (n = 90) M (SD)	Medium (n = 60) M (SD)	High (n = 90) M (SD)	Low (n = 30) M (SD)	Medium (n = 30) M (SD)	High (n = 30) M (SD)	Medium (n = 30) M (SD)
Actual achievement in percent	19.16 (8.62)	52.93 (9.80)	84.01 (8.05)	18.48 (9.84)	50.80 (11.57)	84.80 (6.42)	58.68 (11.25)
Achievement judgment in percent	39.70 (18.68)	54.83 (19.65)	73.89 (17.96)	30.90 (22.65)	56.43 (19.13)	81.17 (14.85)	64.90 (16.03)
Number of calls	16.98 (7.50)	16.03 (6.61)	15.83 (7.18)	16.87 (7.51)	16.57 (7.49)	15.77 (7.84)	16.37 (7.40)

Note. The n indicates the number of available judgments for different types of students, categorized by ethnicity and achievement level. Note that students' actual achievement levels slightly differ from the predetermined achievement level (20%, 50%, 80%) as a probability algorithm was implemented.

Table 4
Unstandardized and Standardized Results of the Moderation Analysis to Predict Judgments of Student Achievement (Given on the Percentage Scale) in Study 2

Predictor	*B*	*SE*	β	*SE*
Intercept	56.16***	1.36		
Achievement (percentage scale)	.53***	.06	.63***	.07
Ethnicity (0 = German/majority, 1 = Asian/minority)	.28	1.82	.01	.07
Interaction Ethnicity × Achievement	.23**	.09	.26**	.10
Number of calls	.06	.13	.02	.04
Interaction Number of Calls × Achievement	−.00	.01	−.03	.06

Note. $R^2 = .502$.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Students' achievement. The achievement parameters were set so that four students had a low achievement (probability of a correct answer: approx. 20%) and four students had a high achievement (probability of a correct answer: approx. 80%). The remaining two students were programmed to show medium achievement (probability of a correct answer: approx. 50%). For half of the sample the German students were among the high-achieving students, for the other half of the sample the German students were among the low-achieving students.

Teachers' judgments. Participants had to judge students' achievement by the percentage of correct answers given by each student on a scale from 0%–100%.

Number of calls. We recorded the number of times each student was called on.

Results

The approach to analyzing the data was analogous to Studies 1 and 2. Table 5 gives an overview of the measures for the different student groups.

To examine the influence of the minority status we took the same approach as in the Studies 1 and 2. We tested the coefficients for the moderation analysis predicting participants' judgments from students' actual achievement and the moderating effect of the minority characteristic on the relationship between judgments and actual achievement (see Table 6).

Belonging to a minority was not associated with biased judgments in one direction, but again was associated with more accurate judgments. As shown by the statistically significant coefficient of the interaction term, we found support for the minority hypothesis: The German students, belonging to the minority in this study, were judged more accurately. Again, we included the number of calls as a further moderator. The analysis showed that the number of calls was not related to different judgments or to the relationship between judgments and actual achievement.

Discussion

In Study 3, we simulated a class in which German students were the minority group and Turkish students formed the majority group. In this way, we wanted to test the minority hypothesis in contrast to the ethnicity hypothesis. Once more, students' minority status moderated the relationship between judgments and actual achievement. German students—constituting the minority—were judged more accurately. This finding confirms the findings of Studies 1 and 2. It further indicates that greater judgment accuracy for a minority group does not seem to be linked to ethnicity but to minority status.

A minority status in a group may arise due to several characteristics—a person's ethnicity, gender, or other attributes (e.g., Heilman, 1980; Kanter, 1977; McArthur & Solomon, 1978; Taylor & Fiske, 1978). Members of a minority are salient, perceived with greater attention, and scrutinized in more detail (Kanter, 1977). To further validate these findings, a minority characteristic other than ethnicity needs to be considered. Within the described studies above, where judgment accuracy was examined, we varied students' ethnicity and minority/majority status and were able to show that—independent of the students' ethnicity—students belonging to a minority group were judged more accurately. In order to validate these findings, we wanted to test the minority hypothesis for another student characteristic, namely student gender.

Table 5
Measures in Study 3

Minority status	Majority (Turkish)			Minority (German)	
Achievement level	Low ($n = 144$) *M* (*SD*)	Medium ($n = 96$) *M* (*SD*)	High ($n = 144$) *M* (*SD*)	Low ($n = 48$) *M* (*SD*)	High ($n = 48$) *M* (*SD*)
Actual achievement in percent	19.52 (5.23)	51.52 (5.55)	83.63 (5.31)	19.90 (5.20)	83.34 (4.61)
Achievement judgment in percent	34.03 (19.47)	56.84 (15.18)	77.33 (15.87)	26.21 (17.80)	81.98 (11.40)
Number of calls	26.44 (9.58)	26.10 (8.79)	25.40 (9.74)	28.23 (12.07)	22.85 (8.59)

Note. The *n* indicates the number of available judgments for different types of students, categorized by minority status and achievement level. Note that students' actual achievement levels slightly differ from the predetermined achievement level (20%, 50%, 80%) as a probability algorithm was implemented.

Table 6
Unstandardized and Standardized Results of the Moderation Analysis to Predict Judgments of Student Achievement (Given on the Percentage Scale) in Study 3

Predictor	*B*	*SE*	β	*SE*
Intercept	55.92***	1.13		
Achievement (percentage scale)	.68***	.04	.74***	.05
Minority status (0 = Turkish/majority, 1 = German/minority)	−2.09	2.05	−.08	.08
Interaction Minority Status × Achievement	.20**	.06	.22**	.06
Number of calls	.06	.11	.02	.04
Interaction Number of Calls × Achievement	−.00	.00	−.03	.04

Note. $R^2 = .634$.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Study 4

In Study 4, we chose gender as the characteristic constituting minority and majority groups of students. Concerning the gender of students, there are also studies on teacher expectations as well as on teacher judgment accuracy. In their meta-analysis on teacher expectancy effects, Dusek and Joseph (1983) reviewed 16 studies in which student gender was related to a measure of teachers' expectations of students' academic performance. The results indicated that students' gender does not affect teachers' expectations of general academic achievement. Jussim and Eccles (1995) reported that teachers perceived girls in fifth grade mathematics classes as performing slightly higher than boys, but this perception was accurate as girls in their study slightly outperformed boys on standardized achievement tests. Ready and Wright (2011) reported an overestimation of girls' literacy skills, but attributed this finding partly on the large sample size of girls with high literacy skills in the study. Overall, there seem to be no differences in teachers' expectations and teachers' judgment accuracy for student gender as an individual student characteristic.

We simulated a class of 16 German students and varied students' gender systematically to form different minority groups. The participants were randomly assigned either to classes with a majority of boys and a minority of girls (boys' classes), or to classes consisting mainly of girls, with boys forming the minority (girls' classes). The rest of the experimental design did not differ from the previously described Studies 1 to 3. As we did not expect the participants to have different expectations for girls' and boys' achievement, we expected to find greater judgment accuracy for the (respective) minority students and thus to find further validation for the minority hypothesis.

Sample

The sample comprised $N = 52$ teacher candidates (65.4% female) at a German university. Participants were being trained to teach in secondary schools and were studying various subject combinations. In Germany, teacher candidates study (at least) two subjects as preparation for their service. Most candidates (94.3%) studied at least one language. In second place, at least one of the subjects studied was a humanities or social sciences subject (57.7%). Followed by science or mathematics (50.0%) and 11.6% were studying at least one subject that was sports, arts, or music. The candidates' mean age was 23.8 ($SD = 2.9$) years. On average, they were in their seventh semester of study ($M = 6.9$, $SD = 3.2$).

Measures

Students' minority status. For roughly half of the sample, the simulated class consisted of 12 boys and four girls, therefore girls were the minority. For the other half of the sample, the simulated class consisted of 12 girls and four boys, thus boys were the minority.

Students' achievement. The students were assigned to a probability of 20%, 40%, 60%, or 80% for giving a correct answer in groups of four (three boys and one girl or three girls and one boy, respectively).

Teachers' judgments. Again, participants were asked to rate the percentage of correct answers given by each student on a scale from 0%–100%.

Number of calls. We recorded the number of calls for each student.

Results

Table 7 presents means and standard deviations for the students' actual achievement, the achievement judgment, and the number of calls for the minority and majority students.

As in the previous studies, we ran a moderation analysis in order to answer the question of whether teachers' judgments were more accurate for minority students. In the moderation analysis, students' actual achievement, students' minority status, and the interaction of minority status with actual achievement were included to predict teacher candidates' judgments. We also included student gender and the interaction of minority status and student gender as further predictors. As shown in Table 8, student gender as a minority characteristic alone was not a significant predictor. The results indicated that students' gender as a minority characteristic moderated the relationship between judgments and students' achievement. This supported the minority hypothesis: The minority with regard to gender was judged more accurately than the majority. Furthermore, we did not find different judgments for boys and girls. We did not find a statistically significant effect for the interaction of minority status and student gender. Thus, it made no difference whether boys or girls were in the minority. We also included the number of calls and the interaction of the number of calls with actual achievement as predictors in the regression analysis in order to see whether participants' calling on behavior moderated the relationship between judgments and students' achievement. We found a statistically significant interaction term

Table 7
Measures In Study 4

Minority status	Majority				Minority			
	Low			High	Low			High
Achievement level	20% (n = 156) *M (SD)*	40% (n = 156) *M (SD)*	60% (n = 156) *M (SD)*	80% (n = 156) *M (SD)*	20% (n = 52) *M (SD)*	40% (n = 52) *M (SD)*	60% (n = 52) *M (SD)*	80% (n = 52) *M (SD)*
Actual achievement in percent	18.21 (8.85)	40.36 (11.41)	63.85 (8.52)	86.06 (7.72)	19.02 (7.86)	41.96 (9.30)	65.16 (7.77)	85.69 (7.96)
Achievement judgment in percent	42.94 (19.03)	52.69 (16.28)	60.65 (17.39)	68.38 (18.72)	37.13 (20.33)	46.48 (20.53)	62.31 (19.30)	74.60 (16.83)
Number of calls	15.30 (7.66)	16.47 (8.19)	16.24 (8.90)	14.48 (7.28)	15.27 (5.99)	16.50 (8.27)	17.14 (7.99)	15.36 (6.47)

Note. The n indicates the number of available judgments for different types of students, categorized by minority status and achievement level. Note that students' actual achievement levels slightly differ from the predetermined achievement level (20%, 40%, 60%, 80%) as a probability algorithm was implemented.

indicating that the higher the number of calls was, the higher the relationship between judgments and students' achievement.

Discussion

In contrast to Studies 1 to 3 where we varied students' ethnicity in the Simulated Classroom, we varied another student characteristic in Study 4, namely student gender. This characteristic may constitute a majority or a minority group with regard to a class distribution. To test the minority hypothesis, we examined whether a group of girls or boys, representing the minority in a class, was judged more accurately than the majority. Concerning the question of whether judgments are biased toward minority students, the results extended the findings of Studies 1 to 3 and validated the minority hypothesis. Again, a difference in judgment accuracy was found for minority students compared to majority students. Whichever gender group represented the minority of their class, members of this minority were judged more accurately.

General Discussion

The present article presented four experimental studies that explored teachers' judgment accuracy when judging minority students' achievement. Our research aim was to examine whether teachers' judgments (Research Question 1) and teachers' judgment accuracy (Research Question 2) were influenced by students' ethnicity and their minority status. Because students' ethnicity and gender can be confounded with students' achievement and, in the field, it is almost impossible to systematically vary students' minority status and control for confounding variables like socioeconomic status, we applied an experimental setting, the Simulated Classroom, in all four studies. This instrument had already been utilized to study teachers' judgment accuracy and proved to yield results comparable to those obtained in field studies (Kaiser et al., 2013; Südkamp & Möller, 2009). More importantly, the Simulated Classroom allowed us to disentangle students' characteristics and minority status and to systematically vary achievement levels between groups.

In Studies 1 and 2 we examined Research Question 1. Based on the meta-analysis by Tenenbaum and Ruck (2007) on teacher expectations, we hypothesized to find a bias of lower (and respectively higher) achievement judgments according to the dominant expectation for that particular group of ethnic minority students, while controlling for students' achievement (expectation hypothesis). We did not find support for this hypothesis. Thus, there was no bias in teachers' judgments of ethnic minority students' academic achievement. We triggered positive and negative expectations and found the same result: Expectations related to ethnicity did not play a role in teachers' judgments, thereby supporting the results of Baker et al. (2015), Hachfeld et al. (2010), and Madon et al. (1998). We take this as a positive outcome. Furthermore, our results are in line with findings from Glock and Krolak-Schwerdt (2014). The authors conducted an experimental study to investigate the information processing of teachers and whether teachers would activate and apply categorical knowledge during their judg-

Table 8
Unstandardized and Standardized Results of the Moderation Analysis to Predict Judgments of Student Achievement (Given on the Percentage Scale) in Study 4

Predictor	*B*	*SE*	β	*SE*
Intercept	57.06***	1.80		
Achievement (percentage scale)	.39***	.04	.49***	.05
Minority status (0 = majority, 1 = minority)	−2.28	2.60	−.11	.12
Interaction Minority Status × Achievement	.20**	.04	.25***	.05
Student gender (0 = boys, 1 = girls)	−1.14	2.32	−.05	.11
Interaction Minority Status × Student Gender	1.50	3.80	.07	.18
Number of calls	−.02	.08	−.01	.03
Interaction Number of Calls × Achievement	.02***	.00	.19***	.04

Note. R^2 = .308.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

ment process. Their participants read a description of a student's behavior in the classroom and received the questions and answers to a German and a mathematics test. Half their participants were also presented with stereotype-activating information that this student was Turkish. Participants who received this additional information about migration background remembered more information from the description and also retrieved more information that was not part of the given information compared to participants whose described student did not have a migration background. But judgments about the students' learning habits, language proficiency, intellectual power, performance in German, and performance in mathematics did not differ between participants who judged a Turkish or a German student. Judgment accuracy was not a part of this study. Glock and Krolak-Schwerdt (2014) reasoned that categorical knowledge does not necessarily influence all stages of information processing. It seems that teachers are able to consciously control the impact categorical knowledge might have on their judgments. Nevertheless, measures of teachers' attitudes about ethnic minority students should also be included in future studies on teacher judgment accuracy.

As expressed in Research Question 2, we further wanted to know whether there were differences in teachers' judgment accuracy for (ethnic) minority students compared to majority students. In Studies 1 and 2, we did this by postulating an ethnicity hypothesis (ethnic minority students are judged differently since they are more salient for teachers). According to this hypothesis our results showed: Ethnic minority students were judged more accurately, regardless of their achievement being high or low. However, we did not disentangle students' ethnicity and minority status. Therefore, in Studies 3 and 4 we tested the minority hypothesis, postulating that not the ethnicity but the minority status as such led to more accurate teacher judgments. In Study 3, we used German students as a minority, and in Study 4, we took students' gender to form a minority group. The results showed support for the minority hypothesis: All four studies revealed greater teacher judgment accuracy for students forming a minority in their class. This was true when ethnicity was used as the minority characteristic (Turkish, Asian, and German in Studies 1, 2, and 3, respectively) as well as when gender was the minority characteristic (in Study 4).

The explanation of the more accurate judgments of students in a minority group could be derived from the findings described by Fiske and Taylor (1991). They assume that minorities are salient and discuss the idea that there is a better memory for salient persons, that they get more attention, that there is better information processing in relation to them and that the information retrieval about them is facilitated. With our design we were able to collect data on teachers' search for information in terms of the number of calls for each student (assuming that the search for information is a behavioral proxy for attention). Apart from Study 4, we did not find a moderating effect of more information (in terms of the number of times a student was called on) on judgment accuracy. However, in Study 4, we found a moderating effect of the number of calls on teachers' judgment accuracy. Students who were called on more often were judged more accurately. This effect occurred in addition to the minority students' moderating effect on judgment accuracy. This finding raises more questions, as Study 4 was the only study in which we implemented gender as the minority characteristic and where we simulated a class of 16 students. One possible explanation could be that participants had more difficulties with the higher number of students. Another could be that students' gender was not as distinct a minority characteristic as students' ethnicity. Therefore, further studies would have to systematically vary class size and gender. To examine the question of whether or not minority students get more attention, studies using eye-tracking to determine the time a student is looked at could provide deeper insights into the matter of attention allocation. Thus, it would be possible to examine how attention moderates judgment processes and under what circumstances the salience of minority students leads to more attention. Higgins (1989) suggested another possibility for how salience could influence judgments. Salient stimuli do not receive much more attention than nonsalient stimuli but salient stimuli receive sufficient attention permitting judgments of them, whereas nonsalient stimuli do not (see also Taylor & Thompson, 1982). For our studies this argument needs to be slightly modified because our participants were explicitly asked to judge salient and nonsalient students. It is difficult to find a definition of when judgments are inaccurate. Rather, there were comparisons between the deviations of judgments and achievement leading to more accurate judgments than inaccurate judgments. Thus, salient students received sufficient attention permitting more accurate judgments. To further examine whether a better memory for minority students could explain the effects that we found, future studies could include further measures testing the memory. For example, we could ask the participants which of their students they still remembered.

Our findings corroborate the importance of considering the classroom level when examining teachers' judgment accuracy. An individual student's characteristic needs to be seen in the context of the class the student is in, because the characteristic might have an additional impact on teachers' judgments in this context. Südkamp and Möller (2009) have already shown a reference-group-effect of the mean achievement level of the class on teachers' judgment accuracy. Ready and Wright (2011) considered the classroom as well as the school context in their study on teachers' judgments. They found a positive association between students' SES and teachers' judgments, which was even stronger for the classroom level. Students' SES influenced teachers' judgments more in classrooms with a generally low SES compared to classrooms with a generally high SES. Ready and Wright also explain the effect of a stronger association between students' SES and teachers' assessments for the classroom level with the students' salience. In a high-SES classroom a child with a low SES is more likely to stand out. Further studies on teachers' judgment accuracy and the moderating effect of classroom composition could examine the role of the class size or the variance in students' achievement.

Another point that needs to be discussed is whether we should interpret our findings in the sense of more accurate or more extreme judgments for minorities. The tables with the descriptive results of actual achievement and judgments show that minority students are judged more extremely compared with majority students. According to some studies (Eisen & McArthur, 1979; Heilman, 1980; McArthur & Solomon, 1978), a person with a minority characteristic is judged more extremely than a person without such a characteristic. Kanter (1977) described the phenomenon that differences to the majority are exaggerated. In our studies, more extreme judgments also meant more accurate judgments. However, we would argue that the explanation of more accurate

judgments is true, as participants did not show exaggerated judgments in the direction of overestimating high-achieving and underestimating low-achieving students, but merely underestimating high-achieving minority students less than high-achieving majority students and overestimating low-achieving minority students less than low-achieving majority students. To further explore this matter, the central tendency in the judgments needs to be reduced. One possible future study with a lower achievement range, for example, probabilities for correct answers between 40% and 60%, could provide further insight. Thereby, the erroneous central tendency should be diminished so that exaggerated judgments could be distinguished from accurate judgments.

We used correlational analyses rather than difference scores to be able to compare our results with existing research in the field of teacher judgment accuracy. With our moderation approach we were able to present measures of correspondence between students' achievement and teachers' judgments as well as whether there were differences in teachers' judgment levels for minority and majority students while controlling for actual achievement.

Differences Between Conditions in the Real Classroom and the Simulated Classroom

Using the Simulated Classroom, we implemented an experimental approach to study teachers' judgment accuracy with high internal validity. Correspondingly, our results cannot be generalized directly to the real classroom situation. Wang, Treat, and Brownell (2008) noted the limited generalizability of studies that emphasize internal validity but highlight the importance of such studies that are able to control for many variables and thus enhance the interpretability of findings. We chose one teaching scenario that does not give justice to the diverse situations teachers are faced with in their daily lives. However, we were able to create an experimental setting that allowed us to present a section of a teaching process that includes the important variables for our research questions without interference of further confounding variables.

In our experimental design we chose a typical instructional scenario with a teacher asking questions and students answering them. In contrast to real classrooms, teacher–student interaction was limited to question-answer-sequences in the Simulated Classroom. On the one hand this means that we were able to control for any influences that might stem from knowledge about student characteristics gained from previous interactions. On the other hand we simplified the possibilities of teacher-student interactions and restricted the authenticity of the teacher–student relationships. The interactions were limited to a very short period of time in the Simulated Classroom and there were also no interactions between students. Whereas in real classrooms teachers have a lot of time to get to know their students, interact with their students in various ways and can also observe interactions between students. As compared with real classroom situations, in the Simulated Classroom participants ask numerous questions in a row within a relatively short amount of time. They can also pose the same question again and again.

Although experiments create artificial, isolated conditions and therefore portray just a fragment of teaching processes, it "mirrors reality" (Taylor & Fiske, 1978; p. 252). Furthermore, our findings are validated by the fact that we found the same pattern of results for different samples. We mainly had experienced teachers as participants, but we were also able to show the same effects with education students and teacher candidates.

Although we were able to ensure high internal validity, it was at the expense of ecological validity. Our teachers could interact with the simulated students only by asking predetermined questions and the students were all preset equally to volunteer to answer questions. In real-life classrooms, engagement and interaction varies between students and low-achieving students especially are often less engaged and teachers do not interact with them as much as with high-achieving students (Mitman, 1985). As the focus of our studies was on disentangling student characteristics we did not consider different kinds of interactions. Although our four studies provided similar results throughout slightly different settings, it remains questionable as to what extent the same findings can be expected in real-life classrooms. Furthermore, the number of students used in our experimental setting ranged from 10 to 16 and this also resulted in a varying number of minority students. In real-life classrooms, teachers have to deal with more students in their classes. We chose this comparably small number of students due to the short time we allowed for interaction and bearing in mind that our participants were not able to make notes on students' achievement. We hoped to keep the cognitive load at a manageable level. Although one could argue that in real life classrooms more students need to be monitored and hence the cognitive load should be even higher we would argue that the disadvantage of a higher number of students is counterbalanced by more time and possibilities to interact with students. This assumption is corroborated by a study investigating the effect of class size on teachers' judgment accuracy. Wild and Rost (1995) did not find a connection between the number of students and more accurate judgments. Besides that, the number of minority students also varied during the four studies. In our opinion, this change in experimental design can be seen as a strength rather than as a limitation, due to the fact that the findings are similar across different settings.

We will use the heuristic model by Südkamp et al. (2012) to further explain the differences between conditions in the real classroom and the conditions in our design as provided in the Simulated Classroom. Concerning student characteristics, the Simulated Classroom makes it possible to vary student characteristics, and thus complements studies conducted in the field. Especially because only a few moderators of teachers' judgment accuracy are known, it is important to control for as many variables as possible to be able to isolate the effects of each characteristic. Compared with real classroom conditions where manifold student characteristics are present we reduced student characteristics to achievement, engagement, gender, and ethnicity in our design, allowing us to examine causal effects of these student characteristics. It also allowed us to establish salient characteristics of students comprising the minority versus the majority in the classroom. In real life classrooms, students' minority or majority status may be related to different student characteristics. A student may even belong to the minority concerning one characteristic (e.g., ethnicity) and he or she may belong to the majority concerning another characteristic (e.g., gender).

Taking the findings of the meta-analysis of teachers' judgment accuracy (Südkamp et al., 2012) into account, test and judgment characteristics also need to be considered, as they are associated with teacher judgment accuracy. Two moderating effects were

found in the meta-analysis. First, informed teacher judgments (compared with not being informed about the achievement test to which the judgment would be related) led to a higher correlation between teachers' judgments and students' academic achievement. Second, higher congruence between the teachers' rating task and the achievement test administered to students was related to greater judgment accuracy.

In the real classroom, there is a multitude of possible teacher judgments depending on the degree of formalization and they have more or less far-reaching consequences. Some judgments, for example for ongoing instruction deciding whether a majority of the class understood the task, are "on the fly" and rather implicit judgments. Given grades on the other hand are explicit judgments and can decide academic opportunities (see Ferreira, Garcia-Marques, Sherman, & Sherman, 2006). Test characteristics can be as numerous as judgment characteristics in the real classroom.

In the Simulated Classroom, teacher judgments were made on a percentage scale. Asking participants for informed judgments creates maximal congruence between the teachers' rating task and the students' achievement measure, thereby controlling for the two moderators of teachers' judgment accuracy designated by Südkamp et al. (2012). Furthermore, we eliminated the random error in the measure of student achievement. In field studies, teacher judgments and students' achievement measures both contain random error. Correlating two measures with random error will result in a lower absolute correlation compared with a removal of the random error from one of those measures. Thus, the design used in our studies produces less random error that could obscure systematic moderating effects. Psychological phenomena in the context of teachers' judgments that have been found in the field can be investigated more minutely under experimental conditions (Wang et al., 2008).

Referring to teacher characteristics, our experimental design also led to differences compared with real classroom conditions. In the Simulated Classroom, teachers' information processing is in the focus. They are asked to select questions, select students to answer the questions, keep track of students' responses, and integrate the information to a judgment. Instructional situations in real classrooms are more complex. Here, for example, teachers additionally need to consider the difficulty of tasks for individual students (pedagogical content knowledge; see Shulman, 1987) and they need knowledge about correct and incorrect solutions to tasks (content knowledge). Whereas teachers in real classrooms need content knowledge and pedagogical content knowledge as well as pedagogical knowledge (for a definition of pedagogical knowledge see Voss et al., 2011) for their judgments we created a situation where only pedagogical knowledge was necessary.

Strengths, Limitations, and Educational Implications

Our findings contribute to the research on teachers' expectations and judgment accuracy in regard to the achievement of students of different ethnicities. Jussim and Harber (2005) asked for more studies dealing with the accuracy of teachers' judgments associated with students' ethnicity. Meeting this demand, we extended the research of biases toward minority students, including the aspect that students of different ethnicities can also form a minority.

Concerning the generalizability of our results for a subject other than mathematics it remains unclear whether the same pattern of results would be found. In mathematics the answers given by students do not allow a great deal of room for interpretation but can be assessed as either right or wrong. Thus, it could be that subjects requiring student answers where the correctness or incorrectness is less obvious are more vulnerable for expectations and a bias could enter into the judgments. In future studies, it is necessary to consider different subjects to clarify this open question.

One limitation that also needs to be addressed is that it was beyond the scope of our studies to include teacher judgments of future student performances. We did not ask for predictions but asked our participants to evaluate students' achievement on specific tasks. This approach reflects instructional situations where teachers collect moment-to-moment data on students' performance and judge students' performance with close temporal proximity. Nevertheless, asking for both judgments (predictions and postdictions) could have given us the opportunity to study differences in prospective and retrospective teacher judgments and should be considered in future research. As the metacognition literature suggests the timing of the judgment is not trivial (e.g., Leonesio & Nelson, 1990), predictions and postdictions are likely different psychologically and may be influenced by different factors.

We did not have a study where we had no minorities. We chose only ethnicity and gender as minority characteristics and it was beyond the scope of this article to find further characteristics that could define minorities and majorities in a classroom. Our findings need to be validated by further student characteristics that could lead to minority/majority proportions in the classroom. It would be especially interesting to examine how the achievement of students who are in the minority because of their high (or low) achievement is judged.

Furthermore, it needs to be noted as a limitation that we found high standard deviations in teachers' judgments suggesting a high variability in teachers' judgment accuracy. While this reflects a general finding for studies in teachers' judgment accuracy (see Südkamp et al., 2012) it also means that with such high standard deviations it is more difficult to detect differences in teachers' judgments between groups of students. Nevertheless, we found differences in teachers' judgment accuracy for different student groups and thus would assume that our results are valid.

In our four studies, we found evidence that students who represent a minority group in their class are judged more accurately. It needs to be discussed whether the greater accuracy is an advantage or a disadvantage for minority students. High-achieving minority students probably have an advantage as their achievement is seen more positively and could thus lead to a more positive grade compared to majority students with the same achievement. On the other hand, low-achieving minority students could have a disadvantage compared with majority students. But it is possible that special needs of low-achieving minority students could be detected earlier. Further studies on teachers' judgment accuracy concerning student characteristics, like ethnicity or gender, should also take class composition into account. As teachers' judgments are the basis for many instructional (e.g., Alvidrez & Weinstein, 1999) and placement decisions (e.g., Bailey & Drummond, 2006), influence students' self-concepts (e.g., Möller, Pohlmann, Köller, & Marsh, 2009) and can play an important role in students' academic careers (e.g., Harlen, 2005), accurate judgments are clearly desir-

able. Especially for students with learning problems, more accurate judgments might be helpful in supporting their academic career (Hachfeld et al., 2010). However, one needs to keep in mind that representing a minority depends on the context. Maybe in real-life classrooms, a minority status is not established by students' ethnicity or gender but by other characteristics (like social class, clothes, wearing glasses, etc.). Possibly heterogeneous classrooms are preferable as they create various minorities and thus could lead to more accurate judgments.

Nevertheless, accurate judgments should not be limited to a minority and (prospective) teachers should be informed about these effects. In particular, teachers could pay attention to the minority/majority proportions in their classrooms. As the Simulated Classroom proved to be a useful setting for studying judgment processes, it could in future be used to demonstrate to teacher candidates their biases when judging student characteristics and to train their diagnostic skills, without risking negative effects on students.

References

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91,* 731–746. http://dx.doi.org/10.1037/0022-0663.91.4.731

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence in educational assessment of students.* Retrieved from http://buros.org/standards-teacher-competence-educational-assessment-students

Bailey, A. L., & Drummond, K. V. (2006). Who is at risk and why? Teachers' reasons for concern and their understanding and assessment of early literacy. *Educational Assessment, 11,* 149–178. http://dx.doi.org/10.1080/10627197.2006.9652988

Baker, C. N., Tichovolsky, M. H., Kupersmidt, J. B., Voegler-Lee, M. E., & Arnold, D. H. (2015). Teacher (mis)perceptions of preschoolers' academic skills: Predictors and associations with longitudinal outcomes. *Journal of Educational Psychology, 107,* 805–820. http://dx.doi.org/10.1037/edu0000008

Baron, R. M., Tom, D. Y. H., & Cooper, H. M. (1985). Social class, race, and teacher expectations. In J. B. Dusek (Ed.), *Teacher expectancies* (pp. 251–269). Hillsdale, NJ: Erlbaum.

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgments and students' performance across a continuum of rating methods. *School Psychology Quarterly, 23,* 43–55. http://dx.doi.org/10.1037/1045-3830.23.1.43

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology, 85,* 347–356. http://dx.doi.org/10.1037/0022-0663.85.2.347

Chang, D. F., & Sue, S. (2003). The effects of race and problem type on teachers' assessments of student behavior. *Journal of Consulting and Clinical Psychology, 71,* 235–242. http://dx.doi.org/10.1037/0022-006X.71.2.235

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 255–296). New York, NY: Macmillan.

Dusek, J. B., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology, 75,* 327–346. http://dx.doi.org/10.1037/0022-0663.75.3.327

Eisen, S. V., & McArthur, L. Z. (1979). Evaluating and sentencing a defendant as a function of his salience and the perceiver's set. *Personality and Social Psychology Bulletin, 5,* 48–52. http://dx.doi.org/10.1177/014616727900500110

Elliott, J., Lee, S. W., & Tollefson, N. (2001). A reliability and validity study of the dynamic indicators of basic early literacy skills–modified. *School Psychology Review, 30,* 33–49.

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *The Journal of Educational Research, 102,* 453–462. http://dx.doi.org/10.3200/JOER.102.6.453-462

Ferreira, M. B., Garcia-Marques, L., Sherman, S. J., & Sherman, J. W. (2006). Automatic and controlled components of judgment and decision making. *Journal of Personality and Social Psychology, 91,* 797–813. http://dx.doi.org/10.1037/0022-3514.91.5.797

Fiedler, K., Freytag, P., & Unkelbach, C. (2007). Pseudocontingencies in a simulated classroom. *Journal of Personality and Social Psychology, 92,* 665–677. http://dx.doi.org/10.1037/0022-3514.92.4.665

Fiedler, K., Walther, E., Freytag, P., & Plessner, H. (2002). Judgment biases in a simulated classroom–A cognitive-environmental approach. *Organizational Behavior and Human Decision Processes, 88,* 527–561. http://dx.doi.org/10.1006/obhd.2001.2981

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). New York, NY: McGraw-Hill.

Friedrich, R. J. (1982). In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science, 26,* 797–833. http://dx.doi.org/10.2307/2110973

Glock, S., & Krolak-Schwerdt, S. (2014). Stereotype activation versus application: How teachers process and judge information about students from ethnic minorities and with low socioeconomic background. *Social Psychology of Education, 17,* 589–607. http://dx.doi.org/10.1007/s11218-014-9266-6

Hachfeld, A., Anders, Y., Schroeder, S., Stanat, P., & Kunter, M. (2010). Does immigration background matter? How teachers' predictions of students' performance relate to student background. *International Journal of Educational Research, 49,* 78–91. http://dx.doi.org/10.1016/j.ijer.2010.09.002

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20,* 245–270. http://dx.doi.org/10.1080/02671520500193744

Hattie, J. A. C. (2002). Classroom composition and peer effects. *International Journal of Educational Research, 37,* 449–481. http://dx.doi.org/10.1016/S0883-0355(03)00015-6

Heilman, M. E. (1980). The impact of situational factors on personnel decisions concerning women: Varying the sex composition of the applicant pool. *Organizational Behavior and Human Performance, 26,* 386–395. http://dx.doi.org/10.1016/0030-5073(80)90074-4

Helwig, R., Anderson, L., & Tindal, G. (2001). Influence of elementary student gender on teachers' perceptions of mathematics achievement. *The Journal of Educational Research, 95,* 93–102. http://dx.doi.org/10.1080/00220670109596577

Higgins, E. T. (1989). Knowledge accessibility and activation: Subjectivity and suffering from unconscious sources. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 75–123). New York, NY: Guilford Press.

Hoge, R. D. (1983). Psychometric properties of teacher-judgment measures of pupil aptitudes, classroom behaviors, and achievement levels. *The Journal of Special Education, 17,* 401–429. http://dx.doi.org/10.1177/002246698301700404

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59,* 297–313. http://dx.doi.org/10.3102/00346543059003297

Hox, J. J. (2002). *Multilevel analysis techniques and applications.* Mahwah, NJ: Erlbaum.

Hurwitz, J. T., Elliott, S. N., & Braden, J. P. (2007). The influence of test familiarity and student disability status upon teachers' judgments of students' test performance. *School Psychology Quarterly, 22,* 115–144. http://dx.doi.org/10.1037/1045-3830.22.2.115

Jussim, L., & Eccles, J. (1995). Are teacher expectations biased by students' gender, social class, or ethnicity? In Y.-T. Lee, L. J. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 245–271). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10495-010

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9,* 131–155. http://dx.doi.org/10.1207/s15327957pspr0902_3

Kaiser, J., Helm, F., Retelsdorf, J., Südkamp, A., & Möller, J. (2012). Zum Zusammenhang von Intelligenz und Urteilsgenauigkeit bei der Beurteilung von Schülerleistungen im Simulierten Klassenraum [On the relation of intelligence and judgment accuracy in the process of assessing student achievement in the simulated classroom]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 26,* 251–261.

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction, 28,* 73–84. http://dx.doi.org/10.1016/j.learninstruc.2013.06.001

Kanter, R. M. (1977). Some effects of proportions on group life: Skewed sex-ratios and responses to token women. *American Journal of Sociology, 82,* 965–990. http://dx.doi.org/10.1086/226425

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 464–467. http://dx.doi.org/10.1037/0278-7393.16.3.464

Leucht, M., Tiffin-Richards, S., Vock, M., Pant, H. A., & Köller, O. (2012). Diagnostische Kompetenz von Englischlehrkräften bei der Bewertung von Schülerleistungen mit Hilfe des Gemeinsamen Europäischen Referenzrahmens für Sprachen [English teachers' diagnostic skills in judging their students' competencies on the basis of the Common European Framework of Reference]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 44,* 163–177. http://dx.doi.org/10.1026/0049-8637/a000071

Luciak, M. (2003). *The educational situation of migrants and ethnic minorities in 15 EU member states in comparative perspective.* Retrieved from http://www.inst.at/trans/15Nr/08_1/luciak15.htm

Luciak, M. (2006). Minority schooling and intercultural education: A comparison of recent developments in the old and new EU member states. *Intercultural Education, 17,* 73–80. http://dx.doi.org/10.1080/14675980500502370

Madon, S., Jussim, L., Keiper, S., Eccles, J., Smith, A., & Palumbo, P. (1998). The accuracy and power of sex, social class, and ethnic stereotypes: A naturalistic study in person perception. *Personality and Social Psychology Bulletin, 24,* 1304–1318. http://dx.doi.org/10.1177/01461672982412005

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79,* 280–295. http://dx.doi.org/10.1037/0022-0663.79.3.280

McArthur, L. Z., & Solomon, L. K. (1978). Perceptions of an aggressive encounter as a function of the victim's salience and the perceiver's arousal. *Journal of Personality and Social Psychology, 36,* 1278–1290. http://dx.doi.org/10.1037/0022-3514.36.11.1278

McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology, 46,* 235–261. http://dx.doi.org/10.1016/j.jsp.2007.05.001

Mitman, A. L. (1985). Teachers' differential behavior toward higher and lower achieving students and its relation to selected teacher characteristics. *Journal of Educational Psychology, 77,* 149–161. http://dx.doi.org/10.1037/0022-0663.77.2.149

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79,* 1129–1167. http://dx.doi.org/10.3102/0034654309337522

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267–316). Washington, DC: American Sociological Association. http://dx.doi.org/10.2307/271070

Muthén, L. K., & Muthén, B. O. (2010). Mplus (Version 6) [Computer software]. Los Angeles, CA: Author.

National Board for Professional Teaching Standards. (2002). *What teachers should know and be able to do.* Arlington, VA: Author. Retrieved from http://www.nbpts.org/UserFiles/File/what_teachers.pdf

Peterson, E. R., Rubie-Davies, C., Osborne, D., & Sibley, C. (2016). Teachers' explicit expectations and implicit prejudiced attitudes to educational achievement: Relations with student achievement and the ethnic achievement gap. *Learning and Instruction, 42,* 123–140. http://dx.doi.org/10.1016/j.learninstruc.2016.01.010

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R. (2008). *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich [PISA 2006 in Germany].* Münster, Germany: Waxmann.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48,* 335–360. http://dx.doi.org/10.3102/0002831210374874

Roick, T., Gölitz, D., & Hasselhorn, M. (2004). *DEMAT 3+. Deutscher Mathematiktest für dritte Klassen* [DEMAT 3+. German mathematics test for third grades]. Göttingen, Germany: Beltz.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development.* New York, NY: Holt, Rinehart & Winston.

Rubie-Davies, C., Hattie, J., & Hamilton, R. (2006). Expecting the best for students: Teacher expectations and academic outcomes. *The British Journal of Educational Psychology, 76,* 429–444. http://dx.doi.org/10.1348/000709905X53589

Schneider, B., & Lee, Y. (1990). A model for academic success: The school and home environment of East Asian students. *Anthropology & Education Quarterly, 21,* 358–377. http://dx.doi.org/10.1525/aeq.1990.21.4.04x0596x

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57,* 1–22. http://dx.doi.org/10.17763/haer.57.1.j463w79r56455411

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz [Accuracy of teacher judgments on student characteristics and the construct of diagnostic competence]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 19,* 85–95.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104,* 743–762. http://dx.doi.org/10.1037/a0027627

Südkamp, A., & Möller, J. (2009). Referenzgruppeneffekte im Simulierten Klassenraum: Direkte und indirekte Einschätzungen von Schülerleistungen [Reference-group effects in a simulated classroom: Direct and indirect judgments]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 23,* 161–174.

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz [The simulated classroom: An experimental study on diagnostic competence]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 22,* 261–276.

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attribution:

Top of the head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 249–288). New York, NY: Academic Press. http://dx.doi.org/10.1016/S0065-2601(08)60009-X

Taylor, S. E., & Thompson, S. C. (1982). Stalking the elusive "vividness" effect. *Psychological Review, 89,* 155–181. http://dx.doi.org/10.1037/0033-295X.89.2.155

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology, 99,* 253–273. http://dx.doi.org/10.1037/0022-0663.99.2.253

Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., . . . Jesse, D. (2015). Can teachers accurately predict student performance? *Teaching and Teacher Education, 49,* 36–44. http://dx.doi.org/10.1016/j.tate.2015.01.012

Trautwein, U., & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind: Referenzgruppeneffekte bei Übertrittsentscheidungen [When high-achieving classmates put students at a disadvantage: Reference group effects at the transition to secondary school]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 21,* 119–133.

Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Self-esteem, academic self-concept, and achievement: How the learning environment moderates the dynamics of self-concept. *Journal of Personality and Social Psychology, 90,* 334–349. http://dx.doi.org/10.1037/0022-3514.90.2.334

van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal, 47,* 497–527. http://dx.doi.org/10.3102/0002831209353594

van Ewijk, R., & Sleegers, P. (2010). Peer ethnicity and achievement: A meta-analysis into the compositional effect. *School Effectiveness and School Improvement, 21,* 237–265. http://dx.doi.org/10.1080/09243451003612671

Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *Journal of Educational Psychology, 103,* 952–969. http://dx.doi.org/10.1037/a0025125

Wang, S. S., Treat, T. A., & Brownell, K. D. (2008). Cognitive processing about classroom-relevant contexts: Teachers' attention to and utilization of girls' body size, ethnicity, attractiveness, and facial affect. *Journal of Educational Psychology, 100,* 473–489. http://dx.doi.org/10.1037/0022-0663.100.2.473

Weiss, H. (2000). Alte und neue Minderheiten. Zum Einstellungswandel in Österreich [Old and new minorities. On the change in attitude in Austria]. *SWS-Rundschau,* 1/2000, 25–42.

Weiss, K. (2007). Zwischen Vietnam und Deutschland—Die Vietnamesen in Ostdeutschland. [Between Vietnam and Germany—The Vietnamese in East Germany] In K. Weiss & H. Kindelberger (Eds.), *Zuwanderung und Integration in den neuen Bundesländern. Zwischen Transferexistenz und Bildungserfolg* (pp. 72–95). Freiburg, Germany: Lambertus.

Westphal, A., Becker, M., Vock, M., Maaz, K., Neumann, M., & McElvany, N. (in press). The link between teacher-assigned grades and classroom socioeconomic composition: The role of classroom behavior, motivation, and teacher characteristics. *Contemporary Educational Psychology.* Advance online publication.

Wild, K.-P., & Rost, D. H. (1995). Klassengröße und Genauigkeit von Schülerbeurteilungen [Class size and the accuracy of teachers' assessments]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 27,* 78–90.

Woodworth, W. D., & Salzer, R. T. (1971). Black children's speech and teachers' evaluations. *Urban Education, 6,* 167–173. http://dx.doi.org/10.1177/004208597100600205

Received May 29, 2015
Revision received June 20, 2016
Accepted August 8, 2016 ■

Call for Papers
A focused collection of qualitative studies in the psychological sciences: Reasoning and participation in formal and informal learning environments

Journal of Educational Psychology
Guest Editors: Tanner LeBaron Wallace and Eric Kuo

Reasoning and participation are two central topics of education research in the psychological sciences. Understanding the mechanisms that govern thought and reasoning has long been a core enterprise of educational psychology and, over time, more modern views on learning have promoted participation as a key feature for research—either as a facilitator of learning, a practice to be learned, or as an operationalization of learning itself.

We are pleased to announce a focused collection highlighting qualitative studies of reasoning and participation in formal and informal learning environments. By inviting studies incorporating qualitative methods, we aim to complement the experimental and longitudinal statistical research on these topics that is typically published in this journal. We encourage submission of papers focused on the following (or closely related) topics:

- Student reasoning and/or participation in novel learning environments or activities
- The relations between student reasoning, motivation, identity, and participation
- Student perceptions and meaning-making during participatory experiences
- Dynamic models of student reasoning that are grounded in data
- Explanatory accounts for how and why participation is successful (or not)
- Identifying new goals or targeted outcomes for reasoning or participation

We especially welcome qualitative studies that demonstrate the possibilities for unique discovery afforded by inductive analysis of rich data sources (e.g., real-time recordings of student reasoning, participation, discourse, and physical action, students' meaning-making anchored to particular interactions experienced). This collection will highlight the benefits of qualitative methods for extending and deepening theoretical and empirical understandings of reasoning and participation in both formal and informal learning environments.

The deadline for manuscript submissions is **March 1, 2018**. We invite authors to contact the Guest Editors of this collection, Tanner LeBaron Wallace (twallace@pitt.edu) and Eric Kuo (erickuo@pitt.edu), for discussion on how to maximize alignment between their submissions and this focused collection, though it is not required. Please follow both APA guidelines as well as specific submission criteria for the journal. When submitting manuscripts, please also indicate your intent to submit to this focused collection in the required cover letter.

All manuscripts must be submitted electronically at http://www.editorialmanager.com/edu. In the submission portal, please select the article type "Special Section: Reasoning & Participation – Qualitative." For more information on the *Journal of Educational Psychology*, please visit http://www.apa.org/pubs/journals/edu/.

Instructions to Authors

Journal of Educational Psychology

www.apa.org/pubs/journals/edu

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Manuscript preparation. Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (6th ed.). Manuscripts may be copyedited for bias-free language (see pp. 70–77 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see www.apa.org/pubs/journals/edu. **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychologial Bulletin, 139,* 133–151. http://dx.doi.org/10.1037/a0028566

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Gill, M. J., & Sypher, B. D. (2009). Workplace incivility and organizational trust. In P. Lutgen-Sandvik & B. D. Sypher (Eds.), *Destructive organizational communication: Processes, consequences, and constructive ways of organizing* (pp. 53–73). New York, NY: Taylor & Francis.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see p. 34 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied in TIFF or EPS format. APA's policy on publication of color figures is available at http://www.apa.org/pubs/authors/instructions.aspx?item=6.

Publication policies. APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at www.apa.org/pubs/authors/posting.aspx. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

Masked review policy. The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., "in our previous work, Johnson et al., 1998 reported that . . .." Instead, references to the authors' work should be in third person, e.g., "Johnson et al. (1998) reported that . . .." The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at www.apa.org/ethics/ or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

Permissions. Authors of accepted papers must obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including test materials (or portions thereof), photographs, and other graphic images (including those used as stimuli in experiments). On advice of counsel, APA may decline to publish any image whose copyright status is unknown.

Supplemental materials. APA can place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see www.apa.org/pubs/authors/supp-material.aspx for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

Submission. Authors should submit their manuscripts electronically via the Manuscript Submission Portal at www.apa.org/pubs/journals/edu/index.aspx (follow the link for submission under Instructions to Authors). General correspondence may be addressed to the incoming editorial office at AConley@apa.org.

NEW RELEASES

from the American Psychological Association

APA Handbook of Forensic Neuropsychology

Editor-in-Chief Shane S. Bush
2018. 528 pages. Hardcover.

Series: *APA Handbooks in Psychology®*
List: $199.00 | APA Member/Affiliate: $129.00
ISBN 978-1-4338-2694-8 | Item # 4311532

Woman's Embodied Self

Feminist Perspectives on Identity and Image
Joan C. Chrisler and Ingrid Johnston-Robledo
2018. 367 pages. Hardcover.

Series: *Psychology of Women*
List: $89.95 | APA Member/Affiliate: $44.95
ISBN 978-1-4338-2712-9 | Item # 4318148

A Telepsychology Casebook

Using Technology Ethically and Effectively in Your Professional Practice
Linda F. Campbell, Fred Millán, and Jana N. Martin
2018. 289 pages. Paperback.

List: $59.95 | APA Member/Affiliate: $39.95
ISBN 978-1-4338-2706-8 | Item # 4317443

125 Years of the American Psychological Association

Edited by Wade E. Pickren and Alexandra Rutherford
2018. 625 pages. Hardcover.

List: $69.95 | APA Member/Affiliate: $34.95
ISBN 978-1-4338-2791-4 | Item # 4316182
Available on Amazon Kindle®

APA Handbook of Giftedness and Talent

Editor-in-Chief Steven I. Pfeiffer
2018. 704 pages. Hardcover.

Series: *APA Handbooks in Psychology®*
List: $199.00 | APA Member/Affiliate: $129.00
ISBN 978-1-4338-2696-2 | Item # 4311533

Mindful Sport Performance Enhancement

Mental Training for Athletes and Coaches
Keith A. Kaufman, Carol R. Glass, and Timothy R. Pineau
2018. 431 pages. Hardcover.

List: $89.95 | APA Member/Affiliate: $49.95
ISBN 978-1-4338-2787-7 | Item # 4317456
Available on Amazon Kindle®

Designing and Proposing Your Research Project

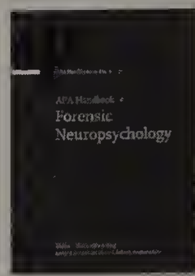
Jennifer Brown Urban and Bradley Matheus Van Eeden-Moorefield
2018. 146 pages. Paperback.

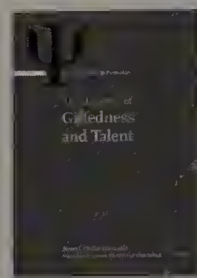
Series: *Concise Guides to Conducting Behavioral, Health, and Social Science Research*
List: $29.95 | APA Member/Affiliate: $25.95
ISBN 978-1-4338-2708-2 | Item # 4313045

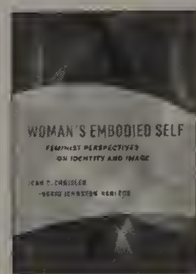
Writing Your Psychology Research Paper

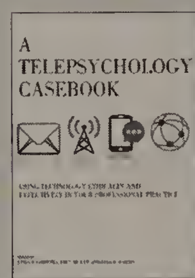
Scott A. Baldwin
2018. 126 pages. Paperback.

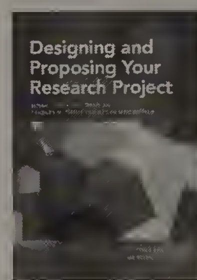
Series: *Concise Guides to Conducting Behavioral, Health, and Social Science Research*
List: $29.95 | APA Member/Affiliate: $25.95
ISBN 978-1-4338-2707-5 | Item # 4313044

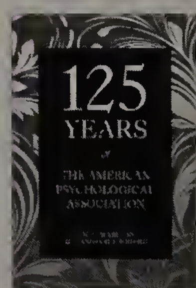


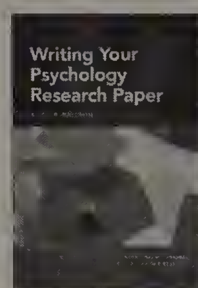












TO ORDER: 800-374-2721 • www.apa.org/pubs/books

AMERICAN PSYCHOLOGICAL ASSOCIATION | 1892 2017

AD3161